

FOUR CONTRIBUTIONS TO
EXPERIMENTAL HEALTH
ECONOMICS

Inauguraldissertation

zur

Erlangung des Doktorgrades

der Wirtschafts- und Sozialwissenschaftlichen Fakultät

der

Universität zu Köln

2021

vorgelegt von

M. Sc. Mona Groß

aus

Düsseldorf

Referent: Prof. Dr. Daniel Wiesen

Korreferent: Prof. Dr. Ludwig Kuntz

Tag der Promotion: 14. September 2021

Acknowledgements

I would like to thank everyone who played a role in my academic accomplishments.

First of all, I would like to express my special thanks to my doctoral supervisor, Prof. Dr. Daniel Wiesen, for his continuous academic and personal support. Second, I am very thankful to Prof. Dr. Ludwig Kuntz for being my second supervisor and for his support during my time at the Department of Business Administration and Health Care Management. I would additionally like to thank Prof. Dr. Andreas Fügener for agreeing to be the Chair of the Doctoral Committee.

I am grateful to all my coauthors for their fruitful input and our inspiring discussions. In particular, I want to thank Dr. Heike Hennig-Schmidt for a listening ear and a helping hand.

My special thanks go to my current and former colleagues at the department for all the moral and professional support. I also want to thank our student assistants for their support related to the projects of my dissertation.

Moreover, I want to thank my family and friends who constantly supported me and filled me with enthusiasm for facing challenges. Finally, I want to thank my late uncle for always showing so much vicarious pride.

Thank you all for your constant personal and academic support!

Cologne, 2021

Mona Groß

Contents

Introduction	1
1 The effects of audits and fines on upcoding in neonatology	6
1.1 Introduction	6
1.2 Experimental design and procedure	7
1.3 Results	10
1.4 Discussion and conclusion	15
2 Physicians' incentives, patients' characteristics, and quality of care: A systematic experimental comparison of fee-for-service, capitation, and pay for performance	18
2.1 Introduction	18
2.2 Contribution to the literature	21
2.3 Experimental design, protocol, and hypotheses	23
2.3.1 Decision situation	23
2.3.2 Payment systems	26
2.3.3 Experimental protocol	27
2.3.4 Behavioral hypotheses	30
2.4 Behavioral results	32
2.4.1 Introductory analyses	33
2.4.2 The effect of blending fee-for-service with performance pay	35
2.4.3 The effect of blending capitation with performance pay	39
2.4.4 Comparison of performance pay effects in capitation and fee-for-service payment systems	42
2.5 Discussion	47
2.5.1 Implications	47
2.5.2 Benefits and costs of introducing performance pay	49
2.5.3 Limitations	51
2.6 Conclusion	54
3 Physicians' performance pay and personality traits	57
3.1 Introduction	57

3.2	Background	59
3.3	Experiment, Data, and Methods	61
3.4	Results	63
3.5	Discussion and conclusion	67
4	Physician altruism: The role of medical education	70
4.1	Introduction	70
4.2	Background and sample	74
4.2.1	Medical education in Germany	74
4.2.2	Our medical student sample	76
4.3	The experiment	77
4.3.1	General design and decision situation	77
4.3.2	Experimental protocol	78
4.3.3	Post-experimental questionnaire	80
4.4	Empirical strategy	81
4.4.1	Behavioral model of altruism	81
4.4.2	Structural estimation	82
4.5	Results	85
4.5.1	Descriptives and non-parametric analyses	85
4.5.2	Structural estimation with observed heterogeneity	88
4.5.3	Robustness of results	91
4.5.4	Patient-regarding altruism, income expectations, and specialty choices	93
4.6	Discussion and conclusion	97
	Bibliography	103
	Appendices	122
A	Appendix to Chapter 1	122
A.1	Additional information about the experiment	122
A.2	Additional analyses	134
B	Appendix to Chapter 2	145
B.1	Additional information about the experiment	145
B.2	Behavioral predictions	163
B.3	Additional analyses	166
C	Appendix to Chapter 3	176
C.1	Additional information about the experiment	176
C.2	Additional analyses	178
D	Appendix to Chapter 4	182
D.1	Additional information about the experiment	182
D.2	Illustration of Delta method	192

D.3	Descriptive analysis of the control group	195
D.4	Further estimations on CES preferences	198
D.5	Alternative behavioral model: Fehr and Schmidt (1999)	221
D.6	Estimation results for patient-regarding altruism, specialty choices, and income expectations	228

List of Figures

1.1	Proportions of reporting behavior by treatments (in %)	11
2.1	Mean quantity by patients' health characteristics	34
2.2	Reduction in the absolute deviation from optimal care by payment system and severity of illness	43
3.1	Distribution of personality traits among subjects by payment conditions	64
3.2	Regression estimates: Effect of performance pay and personality traits on quality of care	65
4.1	Distributions of patient-regarding choices by cohorts	87
4.2	Indifference curves for different study cohorts	89
4.3	Observed vs. unobserved heterogeneity in the random coefficient model	92
A.1.1	Graphical explanation of upcoding	127
A.2.1	Shares of upcoding per subject for treatments NANF and 10AF, differentiated by location	135
B.1.1	Mobile and computer laboratory	145
B.1.2	Patient health benefits by illness and severity of illness	147
B.1.3	Profit parameters in FFS/FFS+P4P and CAP/CAP+P4P	148
B.3.1	Distributions of subjects' quantity choice by severity of illness under different payments scheme	168
B.3.1	Distributions of subjects' quantity choice by severity of illness under different payments scheme (continued)	169
D.3.1	Distributions of patient-regarding choices by cohorts for medical and non-medical students	196
D.4.1	Distributions of parameters a , r and noise for the aggregate model with different sets of covariates, CES preferences	198
D.4.2	Indifference curves for different terms based on random coefficient model, CES preferences	208
D.4.3	Distributions of parameters a , r and noise based on observed heterogeneity for the random coefficient model with different sets of covariates, CES preferences	210

D.4.4	Distributions of parameters a , r and noise based on unobserved heterogeneity for the random coefficient model with different sets of covariates, CES preferences	211
D.4.5	Indifference curves for medical and non medical students for aggregate estimation, CES preferences	213
D.4.6	Distributions of parameters a , r and noise for the aggregate model with different sets of covariates for medical and non medical students, CES preferences	216
D.4.7	Indifference curves for medical and non medical students based on random coefficient model, CES preferences	219
D.5.1	Distributions of parameters a , r and noise based on observed heterogeneity for the random coefficient model with different sets of covariates, Fehr and Schmidt preferences	226
D.5.2	Distributions of parameters a , r and noise based on unobserved heterogeneity for the random coefficient model with different sets of covariates, Fehr and Schmidt preferences	227
D.6.1	Indifference curves for different specialty choices based on random coefficient model, CES preferences	234

List of Tables

1.1	Overview on experimental treatments	8
1.2	Predictive margins from logit regressions on differences in honest reporting between experimental treatments	13
1.3	Predictive margins from a multinomial logit regression on differences in detectable and undetectable upcoding at 1,500g	14
1.4	Overview of mean DRG remunerations per infant by experimental treatment	16
2.1	Experimental parameters	26
2.2	Regression models on the effect on quantity and quality under FFS conditions	37
2.3	Regression models on the effect on quantity and quality under CAP conditions	40
2.4	Comparison of effects of performance pay blended with fee-for-service and capitation on the quality of care	44
2.5	Comparison of effects of blended performance pay systems	46
2.6	Patients' benefits, costs for physicians' remuneration, and changes in costs and benefits	50
3.1	Regression model on the interaction effects of performance pay and personality traits on quality of care	66
4.1	Sample composition by study progress	76
4.2	Our medical student sample in context	77
4.3	Physician profit and patient benefit for treatment alternatives A and B for the 30 patients	79
4.4	Descriptive statistics of medical students' behavior and characteristics . .	86
4.5	Aggregate estimations, preference parameters, noise and marginal effects, CES preferences	90
4.6	Aggregate estimations and random coefficient model, preference parameters, noise and marginal effects, CES preferences, expected income	96
4.7	Aggregate estimations and random coefficient model, preference parameters, noise and marginal effects, CES preferences	98
A.1.1	Profit matrix	126
A.1.2	Decomposition of the effects of audits and fines	126
A.1.3	The 18-items Integrity Scale by Schlenker (2008): Items	133

A.2.1	Comparison of sample characteristics between Bonn and Cologne	134
A.2.2	Individual characteristics by treatment	136
A.2.3	Differences of birth weight reporting by treatments, proportion of participants (in %)	137
A.2.4	Descriptive statistics and analyses of differences in upcoding between treatments (proportion of participants)	138
A.2.5	Descriptive statistics and analyses of differences in detectable and undetectable upcoding at 1,500g (proportions of participants)	139
A.2.6	Treatment effects on the likelihood of detectable upcoding, undetectable upcoding and honest reporting, logit regression models, MEMs	140
A.2.7	Treatment effects on the likelihood of detectable upcoding, undetectable upcoding and honest reporting at true birth weight 1,500g, multinomial logit regression, MEMs	141
A.2.8	Frequency of reported birth weights by true birth weights: No-audit-no-fine treatment	142
A.2.9	Frequency of reported birth weights by true birth weights: 10%-audit-no-fine treatment	143
A.2.10	Frequency of reported birth weights by true birth weights: 10%-audit-and-fine treatment	143
A.2.11	Frequency of reported birth weights by true birth weights: 75%-audit-and-fine treatment	144
B.1.1	Sample characteristics	146
B.1.2	Parameters of main experimental conditions	149
B.3.1	Quantities and qualities of medical service provision by patients' health characteristics and payment system	166
B.3.2	Quantity and quality of health care provision by payment system, illness, and severity of illness	167
B.3.3	Regression models on the effect on quantity and quality between baseline CAP and FFS	170
B.3.4	Regression models on the effect on quantity and quality under FFS conditions without individual control	170
B.3.5	Regression models on the effect on quantity and quality under CAP conditions without individual controls	171
B.3.6	Regression models on the effect on quantity and quality under FFS conditions with the full list of covariates	172
B.3.7	Regression models on the effect on quantity and quality under CAP conditions with the full list of covariates	173
B.3.8	Comparison of effects of blended performance pay systems splitted by marginal health benefit	174

B.3.9	Comparison of effects of blended performance pay systems with the full list of covariates	175
C.1.1	Item description of BFI-10 by Rammstedt et al. (2007)	176
C.1.2	Personality traits by payment systems	177
C.2.3	Regression models on the interaction effects of performance pay and personality traits under CAP, trait by trait analyse	178
C.2.4	Regression models on the interaction effects of performance pay and personality traits under FFS, trait by trait analyses	179
C.2.5	Regression models on several versions of our base model	180
C.2.6	Comparisons of regression models for different regression methods	181
D.1.1	Description of survey items	189
D.3.1	Descriptive statistics of our control group (non-medical students) and our medical student sample	195
D.4.1	Aggregate estimations, parameter estimates, CES preferences	199
D.4.2	Aggregate estimations, preference parameters, noise, and marginal effects, CES preferences	200
D.4.3	Model selection criteria for finite mixture model, CES preferences	203
D.4.4	Finite mixture model, parameter estimates and preference parameters, CES preferences	204
D.4.5	Random coefficient model, parameter estimates, CES preferences	207
D.4.6	Random coefficient model, covariance parameter estimates, CES preferences	208
D.4.7	Random coefficient model, preference parameters, noise, and marginal effects, CES preferences	209
D.4.8	Individual results: descriptive statistics, median and interquartile range, CES preferences	212
D.4.9	Aggregate estimations, parameter estimates for medical and non-medical students, CES preferences	214
D.4.10	Aggregate estimations, preference parameters, noise, and marginal effects for medical and non-medical students, CES preferences	217
D.4.11	Random coefficient model, parameter estimates for medical and non medical students, CES preferences	218
D.4.12	Random coefficient model, covariance parameter estimates for medical and non medical students, CES preferences	219
D.4.13	Random coefficient model, preference parameters, noise, and marginal effects for medical and non medical students, CES preferences	220
D.5.1	Aggregate estimations, preference parameters, noise, and marginal effects, Fehr and Schmidt preferences	223
D.5.2	Random coefficient model, covariance parameter estimates, Fehr and Schmidt preferences	224

D.5.3	Random coefficient model, preference parameters, Fehr and Schmidt preferences	225
D.6.1	Aggregate estimations including income expectations, preference parameters a and r , marginal effects, CES preferences	229
D.6.2	Random coefficient model including income expectations, preference parameters a and r , marginal effects, CES preferences	230
D.6.3	Random coefficient model including income expectations, covariance parameter estimates, CES preferences	231
D.6.4	Descriptive statistics on stated specialty choice	231
D.6.5	Aggregate estimations including specialty choice, preference parameters, noise, and marginal effects, CES preferences	232
D.6.6	Random coefficient model including specialty choice, preference parameters, noise, and marginal effects, CES preferences	233
D.6.7	Random coefficient model including specialty choice, covariance parameter estimates, CES preferences	234

Introduction

“I tried to practice ethical medicine, but it didn’t pay.”

Sandeep Jauhar, “Doctored: The Disillusionment of an American Physician”, 2014

This anecdotal evidence from the memoirs of a practicing US-cardiologist highlights one of the key challenges faced by health care systems around the world: How to incentivize health care service provision to ensure the control of costs and high quality of care at the same time? Rapidly increasing expenditures and mounting inefficiencies in the health care provision (e.g., Baicker and Goldman, 2011; Chandra and Skinner, 2012) reflect the urgent need for action and exert pressure on health care policy-makers to design and implement effective interventions. As these interventions typically address providers, it is of utmost importance to understand the mechanism underlying the responses to policy interventions, such as variations in incentives. Empirical evidence indicates that some interventions to change incentives, which have been promising in theory, either have only moderate effects or lead to unintended consequences when implemented in practice, such as the introduction of performance-based bonus payments (e.g., Scott et al., 2011; Eijkenaar et al., 2013; Mendelson et al., 2017). As one potential cause, the interventions often lack a profound understanding on how they affect individuals’ behavior and what can explain behavioral heterogeneity.

Controlled experiments seem particularly useful in uncovering the behavioral mechanism and can thus complement empirical work in studying the effect of policy interventions (e.g., Falk and Heckman, 2009). The controlled decision environment allows the researcher to exogenously vary regulatory elements such as incentives and thus to causally test whether

and how policy interventions affect individual behavior without confounding factors which are prevalent in the field (e.g., Galizzi and Wiesen, 2017, 2018).

This dissertation consists of four self-contained chapters which explore behavioral responses to policy interventions and preferences of future physicians. The method commonly applied in all four chapters is behavioral experiments in health. Behavioral experiments can represent an indispensable and rather inexpensive “test bed” for gaining insights about behavioral responses to policy interventions and the underlying behavioral mechanism (Galizzi and Wiesen, 2017, 2018). These insights draw a number of lessons to be learned which are valuable to inform health policy debates on the design of interventions before they are rolled out in the field. The first chapter studies how different monitoring policies may influence dishonest behavior in neonatal care. The following two chapters investigate physicians’ responses to financial incentives. In particular, they focus on performance-based bonus payments and the interaction of behavioral responses with patients’ health characteristics (Chapter 2) and physicians’ personality traits (Chapter 3). While it is well established in health economic theory that patient-regarding altruism (Arrow, 1963) in physicians plays a crucial role in their responses to financial incentives (e.g., Ellis and McGuire, 1986), up to now, there is still little empirical evidence on how patient-regarding altruism in future physicians is formed by medical education. The final paper (Chapter 4) addresses this issue and studies patient-regarding altruistic preferences in medical students. Following the taxonomy of Galizzi and Wiesen (2017), the experiments in all chapters are behavioral in the sense that the outcomes of subjects’ decisions consist either of directly observable behavioral responses (Chapters 1-3) or of directly revealed preferences (Chapter 4). In the following, I will briefly outline the chapters of my dissertation.

The first chapter deals with the issue of correcting misaligned financial incentives in the reimbursement of hospital services. Introduced with the intention to reduce and control costs, the change to reimburse hospital services on the basis of diagnosis-related groups unintentionally led to one of the most prevalent types of fraud in health care: upcoding of

patients into a higher reimbursed diagnosis-related group (e.g., Carter et al., 1990; Silverman and Skinner, 2004; Dafny, 2005; Januleviciute et al., 2016; Barros and Braun, 2017; Bastani et al., 2019). To cope with upcoding, monitoring policies, such as audits and fines, are subject to health policy debates. So far, it is not well understood how policies need to be designed to cope with upcoding. With a controlled laboratory experiment I analyze how audits and fines affects upcoding, a behavior which is typically hidden in the field (e.g., Galizzi and Wiesen, 2017, 2018). **Chapter 1** “*The effects of audits and fines on upcoding in neonatology*”¹ provides causal evidence on the effect of random audits with different probabilities and financial consequences on upcoding practices in neonatal care. Using a controlled laboratory experiment, I mimic the decision situation of the obstetrics staff members to report birth weights of neonatal infants. Subjects’ payments in the experiment depend on their reported birth weights and follow the German non-linear diagnosis-related group remuneration for neonatal care. Behavioral results show that audits with low detection probabilities only reduce fraudulent birth-weight reporting, when they are coupled with fines for fraudulent reporting. For audit policies with fines, increasing the probability of an audit only effectively enhances honest reporting, when switching from detectable to less gainful undetectable upcoding is not feasible.

Chapter 2 and 3 address how individuals respond to performance pay and thus contribute to the growing body of literature on the heterogeneity in physicians’ responses to financial performance incentives (e.g., Scott et al., 2011; Mathes et al., 2019, and Jia et al., 2021). **Chapter 2** “*Physicians’ incentives, patients’ characteristics, and quality of care: A systematic experimental comparison of fee-for-service, capitation, and pay for*

¹This paper is co-authored by Hendrik Jürges and Daniel Wiesen. I was responsible for collecting the data, conducted the statistical analyses, and wrote the initial draft. Hendrik Jürges gave input on the econometric specifications and commented on the draft. Daniel Wiesen gave methodological advice and revised the draft. This paper is published and should be cited as follows: Groß, M., Jürges, H. and Wiesen, D. (2021), The effects of audits and fines on upcoding in neonatology. *Health Economics*. 30:1978-1986. doi:10.1002/hec.4272. The layout has been adjusted for the purpose of this dissertation.

performance”² systematically studies how performance pay, complementing either fee-for-service or capitation, affects physicians’ medical service provision and the quality of care. This chapter presents causal evidence from a series of controlled experiments with physicians, medical students, and non-medical students about the effects of performance pay on the quantity and quality of care with heterogeneous patient types. At a within-subject level, performance pay is introduced to complement either fee-for-service or capitation, allowing for between-subject comparisons of the two blended pay for performance systems. A discrete bonus is granted if a quality threshold is reached, which varies with the patients’ severity of illness. Behavioral data show that performance pay significantly reduces non-optimal service provision under fee-for-service and capitation and enhances the quality of care. The effect sizes depend on the patients’ severity of illness. The effect of performance pay and fee-for-service on the quality of care decreases in the patients’ severity of illness, while it increases in severities for performance pay and capitation.

Chapter 3 “*Physician performance pay and personality traits*” studies another potential source of the heterogeneity in behavioral responses to financial incentives, namely physicians’ personality. It explores how responses to performance incentives for physicians aimed at improving the quality of care relate to an individual’s personality traits. This analysis uses the experimental data introduced in Chapter 2. Beyond the experimental evidence that performance pay significantly improves the quality of care under fee-for-service and capitation payment systems, I find differences in the provided quality of care for some personality traits. More conscientious and more agreeable individuals respond significantly less strongly to incentives under performance pay blended with capitation. Other personality traits, such as extraversion, openness, and neuroticism, are not significantly related to individuals’ behavior. Under fee-for-service payments, however, personality traits are not

²This chapter is based on a joint paper with Jeannette Brosig-Koch, Heike Hennig-Schmidt, Nadja Kairies-Schwarz and Daniel Wiesen entitled “Physicians’ incentives, patients’ characteristics, and quality of care: A systematic experimental comparison of fee-for-service, capitation, and pay for performance”. The co-authors developed the experimental design and conducted the experiments. I analyzed the data and wrote the draft which is a completely revised and extended version of a working paper by the co-authors, circulated under the title “Physician performance pay: Evidence from a laboratory experiment” (Ruhr Economic Paper No. 658). Heike Hennig-Schmidt and Daniel Wiesen gave methodological advice and revised the draft. Jeannette Brosig-Koch and Nadja Kairies-Schwarz gave input on the modified direction of the paper and commented on the revised draft.

behaviorally relevant. These findings can be informative for incentivizing physicians better and sorting them into incentive schemes.

Understanding how physicians respond to incentives is important for policy-makers and researchers alike. One particularly important aspect therein is the role of physician altruism (e.g., Arrow, 1963). Physicians' altruistic preferences are not only key in describing responses to financial incentives (e.g., Ellis and McGuire, 1986; Alexander, 2020) but represent a key determinant for physicians' behavior in general (e.g., Arrow, 1963; Allard et al., 2011; Kolstad, 2013). Deeply rooted in the profession, physician altruism is still maintained by the modern-day physician's pledge stating: "The health and well-being of my patient will be my first consideration" (World Medical Association, 2018). Despite its essential role, surprisingly little is known empirically about how physician altruism is formed by medical education. The final chapter presents a novel experimental choice task to elicit physicians' altruistic preferences towards the patients' health. **Chapter 4** "*Physician altruism: The role of medical education*"³ studies how patient-regarding altruism is affected by medical education and presents structural estimates on experimental data from a large sample of German medical students ($N=733$) varying in their study progresses. The estimates reveal substantial heterogeneity in altruistic preferences across study cohorts. Patient-regarding altruism is highest for freshmen, significantly declines for students in the pre-clinical and clinical study phase, and tends to increase for practical-year students who are assisting in clinical practice. Across individuals, patient-regarding altruism is higher for females and increases in general altruism. Altruistic subjects have lower income expectations and are more likely to choose surgery and pediatrics as their preferred specialty.

³This paper is joint work with Arthur E. Attema, Matteo M. Galizzi, Heike Hennig-Schmidt, Yassin Karay, Olivier L'Haridon, and Daniel Wiesen. Daniel Wiesen and Heike Hennig-Schmidt developed the experimental decision task. I was responsible for conducting the experiments, assisted by Heike Hennig-Schmidt, Yassin Karay (and the deanery of the medical faculty), and Daniel Wiesen. I conducted the statistical analyses and prepared the initial draft of the paper. Olivier L'Haridon performed the econometric estimations. Heike Hennig Schmidt, Olivier L'Haridon, Daniel Wiesen, and I jointly wrote the final version of the paper. Arthur E. Attema, Matteo M. Galizzi, and Yassin Karay reviewed and commented the manuscript.

Chapter 1

The effects of audits and fines on upcoding in neonatology

1.1 Introduction

Upcoding of patients to attract higher diagnosis-related group (DRG) payments is a common problem in several healthcare systems, leading to inefficiencies and financial losses (e.g., Carter et al., 1990; Silverman and Skinner, 2004; Dafny, 2005; Januleviciute et al., 2016; Barros and Braun, 2017; Bastani et al., 2019). Incentives to upcode are particularly prevailing in neonatal intensive care (e.g., Shigeoka and Fushimi, 2014; Jürges and Köberlein, 2015; Reif et al., 2018; Hochuli, 2020). The reimbursement for neonatal care is typically determined through birth weights reported by obstetrics staff. DRG payments non-linearly increase with decreasing birth weights at birth-weight thresholds. In Germany, for example, reporting weights just below a threshold may yield additional payments of more than EUR 17,000 (a relative increase of 40%) (Jürges and Köberlein, 2015).⁴

To cope with upcoding, policy-makers often intend to increase the frequencies of audits.⁵

⁴According to the German DRG schedule, reported weights that fall into the category of 750g to 874g instead of 875g to 999g, increases the average reimbursement by EUR 17,555, from EUR 45,985 to EUR 63,540; see Jürges and Köberlein (2015) for further details.

⁵Increasing audit frequencies for a given fine level would make dishonest behavior less attractive (e.g., Becker 1968). In a Beckerian sense, a utility-maximizing decision-maker weighs the expected utilities from dishonest and honest behavior. This logic motivated health-policy reforms that were recently implemented after much debate. For example, in Germany a recent reform of the Medical Service of the Sick Funds (MDK) came into effect in early 2020. It intends to sanction fraudulent reporting of hospital bills. In detail, depending on the overall share of unobjected (or correctly billed) hospital invoices, hospitals must pay a fine to the health insurance funds in addition to the repayment of the difference between the true and the wrongly billed amount. In 2020, the fine amounts to 10% of the difference, but is never less than EUR 300. From 2021, the size of the fine will differ depending on the overall share of contested bills in 2020 (*Gesetz für bessere und abhängige Prüfungen (MDK-Reformgesetz) 2019*, Art. 1, §275c). In the past, German hospitals did not face any consequences beyond the repayment of the falsely billed amounts. Moreover, hospitals have received, and will continue to receive, EUR 300 for every audited invoice which has been correctly billed as a lump-sum expense allowance from the health insurance fund.

However, empirical evidence on the effectiveness of such a means is mostly lacking. Only Hennig-Schmidt et al.’s (2019) experiment with neonatal framing shows that a random audit coupled with a fine effectively reduces dishonesty. Similarly, Angerer et al. (2021) find that non-optimal treatment decreases with more frequent audits in a credence goods experiment.⁶ Nevertheless, it is not well understood whether a random audit alone, which upon detection confronts individuals with their dishonesty alluding to self-image concerns (e.g., Bénabou and Tirole, 2006), is sufficient or whether financial consequences are needed to induce honesty.

Using a behavioral experiment, we investigate how (i) coupling random audits with fines and (ii) increasing the audit probability affects upcoding. A controlled laboratory experiment renders control over the decision situation and allows us to observe upcoding which is typically hidden in the field.⁷ We thus complement field evidence on upcoding in neonatal care by analyzing *individual* behavior in situations in which upcoding is either detectable or undetectable through audits.

1.2 Experimental design and procedure

In a neonatal framing (Hennig-Schmidt et al., 2019), subjects role-play obstetrics staff members charged with the task of entering birth weights \hat{w}_j of six preterm infants j in birth reports. The birth weights w_j are drawn in random order and shown on subjects’ screens: 1,200g, 1,250g, 1,300g, 1,350g, 1,400g, and 1,500g. After having seen an infant’s weight, subjects are asked to report the birth weight $\hat{w}_j = [1,150; \dots; 1,550]$ in 50-gram increments.

Subjects receive a fixed lump-sum payment F and variable DRG-based payments $r(\hat{w}_j^i)$ depending on subject i ’s *reported* weight per infant. Based on empirical evidence (e.g., Reif et al., 2018), subjects are informed that infants receive optimal care according to their *true* (not the reported) birth weights, which excludes non-financial motivations to upcode. Consequently, treatment costs $c(w_j)$ also depend on *true* weights. Subject i ’s overall profit is:

⁶For a definition of credence goods, see, for example, Dulleck and Kerschbamer (2006) and for excellent surveys of the literature, see, for example, Kerschbamer and Sutter (2017) and Balafoutas and Kerschbamer (2020).

⁷For a definition of behavioral experiment in health and more on the discussion of the use of experiments in health economics, see Galizzi and Wiesen (2017, 2018).

$\pi^i(w, \hat{w}) = F + \sum_{j=1}^6 r(\hat{w}_j) - c(w_j)$. For an illustration of a decision situation, see Appendix B.1.4.

The range of weights in the experiment comprises thresholds at 1,250g and 1,500g, following the German DRG-scheme. Payments within a DRG are set such that average treatment costs of an infant are mostly covered. For profits of all combinations of reported and true weights, see Table A.1.1 in Appendix A.1.2. We refer to upcoding whenever a birth weight is fraudulently reported to be below a threshold implying a higher DRG-based payment.

Between-subjects, we test the effects of different audit probabilities and of a fine on

Table 1.1: Overview on experimental treatments

Detection probability	Financial consequences	
	No fine	Fine
0% (no audit)	NANF (No-audit-no-fine/ Baseline): — No random audit, upcoding cannot be detected, no fine ($n= 56$).	
10%	10ANF (10%-audit-no-fine): Random audit of subjects' reported birth weights with 10% probability. If upcoding is detected, subjects are informed about the detection of their fraudulent behavior, but they are not fined ($n= 38$).	10AF (10%-audit-and-fine): Random audit is equivalent to 10ANF. If upcoding is detected, subjects are informed about the detection of their fraudulent behavior, and they are fined. That means they only receive the fixed amount F ($n= 65$).
75%	—	75AF (75%-audit-and-fine): Random audit of subjects' reported birth weights with 75% probability. If upcoding is detected, subjects are informed about the detection of their fraudulent behavior, and they are fined. That means they only receive the fixed amount F ($n= 38$).

Notes. The number of participants per treatment is reported in parentheses.

individuals' reporting behavior; see Table 1.1. In our baseline (treatment NANF), subjects report birth weights without audits and fines. To investigate audit-effects without fines, a 10%-random audit is introduced in 10ANF. A fine for fraudulent reporting is added in 10AF. If upcoding is detected, all DRG-based payments are withheld and subjects only receive the lump-sum F . In 75AF, the probability of an audit (hence detection and fine) is increased

to 75%.⁸ We compare reporting behavior between NANF and 10ANF and between 10AF and 75AF to analyze the *effects of audits*. To analyze the *effect of fines* (at low detection probability), we compare treatments 10ANF and 10AF.⁹

The audit mechanism in 10ANF, 10AF, and 75AF relies on comparisons of reported birth weights with infants' weights recorded on their second day of life. Assuming that hospital records show the correct weight on the second day after birth, and taking into account that newborns lose about 5% of their initial weight within the first 24 hours (e.g., Flaherman et al., 2015), the second-day weight cannot be higher than the first-day weight. The opposite would indicate fraudulent reporting. The second-day weight plus 5% thus represents a lower bound for the true birth weight which enables subjects to upcode without the risk of being detected. The analogy for our experiment is that upcoding by 50g is undetectable and upcoding by more than 100g is detectable. For true birth weights just above the thresholds (1,250g and 1,500g), upcoding to the next DRG-threshold goes undetected. Upcoding 1,500g-infants by two DRG-thresholds is then detectable.¹⁰ In contrast, subjects always face a risk of detection for true birth weights of 1,300g, 1,350g, and 1,400g, for which upcoding is to report 1,200g. We are thus able to investigate effects of audits and fines on honest reporting when either (i) only detectable upcoding is possible, (ii) undetectable upcoding is possible, or (iii) detectable and undetectable upcoding are possible. In case a subject is audited and upcoding is detected (for at least one infant), subjects are informed about the detection of fraudulent behavior.

The computerized experiment programmed in z-Tree (Fischbacher, 2007) was conducted via the Cologne Laboratory for Experimental Research in May and June 2019. For the experimental protocol, see Appendix A.1.5. For treatments NANF and 10AF, we also include data from the experiments by Hennig-Schmidt et al. (2019) which were conducted at the

⁸While the fine seems substantial under 10AF, a risk-neutral profit-maximizer would still upcode weights in all decisions. An audit probability of 75% represents the cut-off value which (assuming common knowledge about randomly drawn birth weights) implies that a profit-maximizer would only engage in undetectable upcoding which cannot be detected by audits.

⁹For a detailed description of the decomposition of the effects of audits and fines, see Appendix A.1.3.

¹⁰In detail, when $w_j = 1,250$ g, reporting a birth weight of $\hat{w}_j = 1,200$ g cannot be detected by an audit, as the reported birth weight is higher than the lower bound of the true weight of 1,187.5g on day two. The other possibility of undetectable upcoding is represented by $w_j = 1,500$ g (lower bound of the true weight at the second weighing: 1,425g). Here, reporting $\hat{w}_j = 1,450$ g cannot be detected by an audit, whereas reporting $w_j = 1,200$ g can. For a graphical explanation of (detectable and undetectable) upcoding, see Figure A.1.1 in the Appendix A.1.4.

BonnEconLab in 2014 and 2015. Characteristics and behavior of subjects from Bonn and Cologne did not differ significantly; see Appendix A.2.1. Overall, 197 students participated in our experiments, 99 in Cologne and 98 in Bonn; 61% were medical students, the average age was 23 years, and 45% were female.

1.3 Results

In all treatments, we observed substantial upcoding of neonatal cases between 37% and 82%. We further differentiate between three types of reporting behavior: detectable upcoding, undetectable upcoding, and honest reporting.¹¹ Figure 1.1 shows the average proportions for the reporting types differentiated by treatments. The observed pattern is that, the introduction of a 10%-random audit without fine (10ANF), the introduction of a fine to the 10%-audit (10AF), and the increase in the audit probability from 10% to 75% (75AF), reduced detectable upcoding while undetectable upcoding increased. Honest reporting increased at a small rate for 10ANF and more substantially for audit policies with fines (10AF, 75AF).

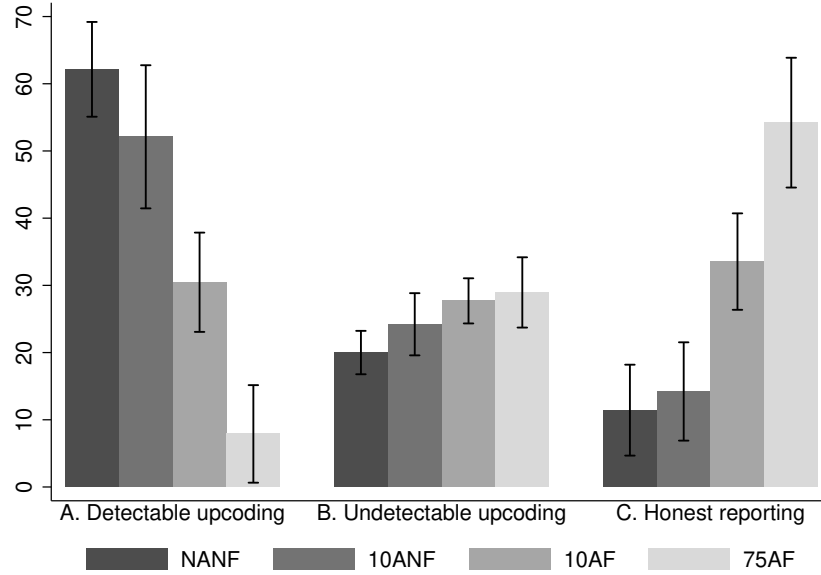
We next separately analyze individuals' honest reporting as opposed to situations in which either only detectable or undetectable upcoding is feasible. We apply logit regression models to analyze individuals' behavior controlling for subjects' characteristics: age, gender, medical major, personality traits (Big-Five Inventory; Rammstedt and John, 2007), and integrity (Schlenker, 2008).¹² Table 1.2 reports predictive margins for honest birth-weight reporting differentiated by treatments. Treatment differences expressed as marginal effects at the means reflect the effects of audits and fines on the likelihood of honest reporting in percentage points (pp).

First, we analyze *honest reporting* only for true birth weights, at which upcoding is always detectable (1,300g, 1,350g, and 1,400g); see Panel A of Table 1.2. Introducing a 10%-audit

¹¹The vast majority of decisions falls in one of the three categories. Only 8% of reported birth weights reports per treatment deviate from these classifications of behavior and are categorized as unclassified. The proportions of unclassified birth weight reports did not vary systematically with the treatment; see Table A.2.3 in Appendix A.2.3. For the frequencies of individuals' choices, see Tables A.2.8 to A.2.11 in Appendix A.2.4.

¹²For descriptive statistics on individual characteristics by treatments, see Table A.2.2 in Appendix A.2.2. Descriptive statistics on proportions in upcoding behavior and reported marginal effects based on regressions without individual controls yield very similar results on the effects of an audit and fine; see Tables A.2.4, A.2.5, A.2.6, and A.2.7 in the Appendix A.2.3.

Figure 1.1: Proportions of reporting behavior by treatments (in %)



Notes. The bar charts represent the average shares of upcoding and honest reporting differentiated by treatments, with 95% confidence intervals. Upcoding is defined as payment-increasing misreports of birth weights. Undetectable upcoding is achieved by misreporting birth weights by only 50g, which cannot be detected by an audit (Panel A). Detectable upcoding implies misreporting birth weights by 100g or more, which can always be detected by audit (Panel B). Honest reporting refers to reporting the true birth weight (Panel C). We included all birth weights except 1,200g where (payment-increasing) upcoding is not possible ($k=985$ decisions).

without a fine slightly increased honest reporting by about 7 pp ($p=0.248$). Adding a fine to the 10%-random audit significantly increased honest reporting by about 33 pp ($p<0.001$). An increase in audit probability from 10% to 75% (with fines) increased honest reporting by roughly 21 pp ($p=0.015$).

Second, we consider the effects of audits and fines on honest reporting at 1,250g, the true birth weight at which upcoding is undetectable; see Panel B of Table 1.2.¹³ Introducing a 10%-audit led to a decrease in honest reporting by about 6 pp ($p=0.194$). Adding a fine to an audit with a low detection probability, increased honest reporting by about 4 pp ($p=0.307$). Raising the audit probability further increased honest reporting by about 16 pp ($p=0.070$).

Finally, we analyze individuals' honest reporting for the true birth weight of 1,500g. At this weight, subjects have the opportunity to choose between detectable and undetectable

¹³Note that at 1,250g, reporting 1,150g would be detectable but it yields no financial gain compared to reporting 1,200g (and remaining undetectable). Participants who choose 1,150g make an inferior decision and we thus classify it "other" reporting behavior.

upcoding where the former yields a higher gain; see Panel C of Table 1.2. Introducing a 10%-audit did not significantly affect honest reporting ($p=0.954$). Adding a fine to it, however, significantly increased honest reporting by about 13 pp ($p=0.023$). Raising the detection probability of a random audit to 75% did also not significantly affect honesty (1 pp, $p=0.887$).

While rational choice options at the previously considered true birth weights are binary, subjects have three choice options at 1,500g: honest reporting, *detectable* upcoding (by two DRG-thresholds) and *undetectable* upcoding (by one threshold). Table 1.3 reports predictive margins based on multinomial logit regressions for detectable and undetectable upcoding differentiated by treatments. Treatment effects are expressed as marginal effects at the means in percentage points (pp). We find that introducing a 10%-random audit decreased detectable upcoding by about 18 pp ($p=0.089$). At the same time, however, undetectable upcoding increased by about 19 pp ($p=0.063$). Obviously, honest reporting was hardly affected. Adding a fine to the 10%-audit significantly decreased detectable upcoding by about 33 pp ($p=0.002$), and in parallel increased undetectable upcoding by about 20 pp ($p=0.065$). With a fine, an increase in audit probabilities from 10% to 75% reduced detectable upcoding by about 17 pp ($p=0.022$) and increased undetectable upcoding by about 16 pp ($p=0.113$). The results reveal an unintended consequence of audits in that they shifted detectable to undetectable upcoding rather than triggering more honest reporting.

In sum, our analyses indicate that, first, random audits with low detection probabilities only reduced upcoding and fostered honest reporting if audits comprise fines. This emphasizes the importance of a financial consequence to cope with dishonest behavior and complements findings from Hennig-Schmidt et al. (2019). Second, raising the probability of an audit increased honest reporting in the decision situations when either only detectable or only undetectable upcoding was possible. As a separate analysis of the true birth weights of 1,500g indicated reductions in detectable upcoding were accompanied by increases in undetectable upcoding.

Table 1.2: Predictive margins from logit regressions on differences in honest reporting between experimental treatments

A. Honest reporting when only detectable upcoding is possible (at 1,300g, 1,350g, and 1,400g; $k=591$ decisions)			
Detection prob.	No fine	Fine	Δ , in pp (p -value)
0% (no audit)	12.4% (NANF)	—	
10%	19.8% (10ANF)	52.4% (10AF)	32.6 (< 0.001)
75%	—	73.8% (75AF)	
Δ , in pp (p -value)	7.4 (0.248)	21.4 (0.015)	
B. Honest reporting when undetectable upcoding is possible (at 1,250g; $k=197$ decisions)			
Detection prob.	No fine	Fine	Δ , in pp (p -value)
0% (no audit)	7.4% (NANF)	—	
10%	1.8% (10ANF)	5.7% (10AF)	3.9 (0.307)
75%	—	21.7% (75AF)	
Δ , in pp (p -value)	−5.6 (0.194)	16.0 (0.070)	
C. Honest reporting when undetectable and detectable upcoding are feasible (at 1,500g; $k=196$ decisions)			
Detection prob.	No fine	Fine	Δ , in pp (p -value)
0% (no audit)	4.3% (NANF)	—	
10%	3.5% (10ANF)	16.3% (10AF)	12.8 (0.023)
75%	—	17.4% (75AF)	
Δ , in pp (p -value)	−0.8 (0.837)	1.2 (0.887)	

Notes. This table reports predictive margins at the means of the covariates based on logit models on honest reporting for Panels A and B and based on a multinomial logit model on reporting behavior at 1,500g for Panel C. Upcoding is defined as payment-increasing misreports of birth weights. Panel A estimates honest reporting for infants for whom upcoding is always detectable in case of an audit (1,300g, 1,350g, and 1,400g; $k=591$ decisions). Panel B estimates honest reporting for infants of 1,250g for whom undetectable upcoding is possible ($k=197$ decisions). Undetectable upcoding means that upcoding is achieved by misreports of birth weights by only 50g, which cannot be detected by an audit. Panel C estimates honest reporting for infants of 1,500g for whom undetectable upcoding is possible but less gainful than detectable upcoding ($k=196$ decisions). All predictive margins are adjusted for individual characteristics, i.e., gender, age, medical major, personality traits (Big-Five Inventory), and integrity (Schlenker, 2008). Treatment effects (Δ) are differences in marginal effects at the means (percentage points, pp). Full regression results are reported in the Appendix, Table A.2.6 and Table A.2.7.

Table 1.3: Predictive margins from a multinomial logit regression on differences in detectable and undetectable upcoding at 1,500g

A. Detectable upcoding			
(at 1,500g; $k=196$ decisions)			
Detection prob.	No fine	Fine	Δ , in pp (p -value)
0% (no audit)	76.2% (NANF)	—	
10%	57.8% (10ANF)	24.9% (10AF)	−32.9 (0.002)
75%	—	7.8% (75AF)	
Δ , in pp (p -value)	−18.4 (0.089)	−17.1 (0.022)	
B. Undetectable upcoding			
(at 1,500g; $k=196$ decisions)			
Detection prob.	No fine	Fine	Δ , in pp (p -value)
0% (no audit)	19.5% (NANF)	—	
10%	38.7% (10ANF)	58.8% (10AF)	20.2 (0.065)
75%	—	74.8% (75AF)	
Δ , in pp (p -value)	19.2 (0.063)	16.0 (0.113)	

Notes. This table reports predictive margins at the means of the covariates based on a multinomial logit model on upcoding at 1,500g ($k=196$ decisions). Upcoding is defined as payment-increasing misreports of birth weights. All predictive margins are adjusted for individual characteristics, i.e., gender, age, medical major, personality traits (Big-Five Inventory), and integrity (Schlenker, 2008). "Detectable upcoding" takes place when an individual reports a weight of 1,250g or lower, "Undetectable upcoding" takes place when an individual reports a (fraudulent birth) weight of 1,450g. Estimates for "Honest reporting" are reported in Table 1.2. We excluded one subject with a birth-weight entry of 1,550g. Treatment effects (Δ) are differences in marginal effects at the means (percentage points, pp). Full regression results are reported in the Appendix, Table A.2.7.

1.4 Discussion and conclusion

Our behavioral experiment provides important insights for healthcare policy-makers on the effects of different audit policies on upcoding in neonatology. First, random audits at low detection probability without financial consequences are not sufficient to foster honest reporting. Only when random audits include fining of fraudulent reporting, honesty increased significantly. A fine seems thus to be an essential instrument carrying a signal that fraudulent reporting of birth weights represents misbehavior and is sanctioned. Second, increasing the frequency of random audits only induces more honesty when individuals are not able to shift from detectable to undetectable upcoding. Hence, differentiating between detectable and undetectable upcoding reveals the unintended consequence of audit policies to foster more undetectable upcoding rather than honest reporting.

When interpreting behavioral consequences for the remuneration within the confines of the experiment, upcoding led to high financial losses for payers. Without an audit, the average payment per infant almost doubled compared to a theoretical payment under fully honest reporting. Introducing a 10%-random audit (without a fine) reduced the financial loss for the insurer by 10%.¹⁴ Only when subjects bore the risk of being fined for fraudulent reporting, however, there was a noticeable drop. When a 10% or 75%-audit came with a fine, the financial loss declined by 36% or 66% compared to no audit; see Panel A of Table 1.4. The effects are more pronounced focusing on infants for whom only detectable upcoding is possible. Introducing a 10%-audit reduced the financial loss by 15%, a 10%-audit and fine by 49%, and a 75%-audit and fine by 87%; see Panel B of Table 1.4.

We now separately consider remuneration effects per infant at 1,250g and 1,500g for whom undetectable upcoding was possible. While treatment comparisons reveal that the insurer's financial loss due to DRG upcoding increased by 4% (10ANF) and by 3% (10AF) for infants at 1,250g, it can be reduced by 17% only when audits occurred with high detection probability and a fine (75AF); see Panel C of Table 1.4. However, the insurer's loss can

¹⁴Note that the reduction in financial loss only refers to the reduced payments the insurer has to pay per infant based on observed reporting behaviors. Potential fines which hospitals have to pay if they are found out for misreporting birth weights are not considered in the calculation and would even lead to higher reductions in financial losses.

Table 1.4: Overview of mean DRG remunerations per infant by experimental treatment

Treatment	Mean remuneration per infant (in Taler)		Loss due to upcoding (in Taler)	Reduction in loss compared to NANF (in %)
	If fully honest	Observed behavior		
A. At all birth weights; $k=985$ decisions				
NANF	184	343	159	—
10ANF	184	327	143	-10.1
10AF	184	286	102	-35.8
75AF	184	238	54	-66.0
B. At 1,300g, 1,350g, and 1,400g; when only detectable upcoding is possible; $k=591$ decisions				
NANF	200	344	144	—
10ANF	200	323	123	-14.6
10AF	200	273	73	-49.3
75AF	200	219	19	-86.8
C. At 1,250g; when undetectable upcoding is possible; $k=197$ decisions				
NANF	200	364	164	—
10ANF	200	371	171	+4.3
10AF	200	369	169	+3.0
75AF	200	337	137	-16.5
D. At 1,500g; when undetectable is possible but less gainful than detectable upcoding; $k=197$ decisions				
NANF	120	320	200	—
10ANF	120	295	175	-12.5
10AF	120	243	123	-38.5
75AF	120	197	77	-61.5

Notes. This table reports the average DRG remuneration the insurer has to pay per infant. Upcoding is defined as payment-increasing misreports of birth weights. In Panel A, we only consider the infants for whom upcoding is possible (birth weight of 1,200g is excluded for our calculations). In Panel B, we only consider the infants for whom upcoding only detectable upcoding is possible (1,300g, 1,350g, and 1,400g). In Panel C, we only consider the infant with birth weight of 1,200g for whom undetectable upcoding is possible. In Panel D, we only consider the infant with birth weight of 1,500g for whom undetectable upcoding is possible but less gainful detectable upcoding. Under full honest reporting, we report the average hypothetical remuneration for the true birth weight of the respective infants. Observed behavior refers to our behavioral data differentiated by treatments. We have calculated the mean remuneration of every subject per infant based on the reported birth weights. We report the financial loss for the insurer due to DRG upcoding as the difference between fully honest reporting and our behavioral data. In the last column, we calculate the relative differences of financial loss between the respective audit and our baseline treatment. All monetary amounts are given in Taler, our experimental currency, the exchange rate being 1 Taler = 0.01 EUR.

be reduced by 13% (10ANF), by 39% (10ANF), and by 62% (75AF) for infants with true birth weights of 1,500g; see Panel D of Table 1.4. This decline can be explained by upcoding infants with true birth weights of 1,500g by one instead of two DRG thresholds. Thus, switching from detectable to less gainful undetectable upcoding reduced the payments to some extent.

The results should be interpreted within the confines of our experimental setup. The implementation of audit policies comes at costs, such as set-up costs for monitoring programs, personnel costs, and costs for potential false positive findings which need to be weighed against the savings in expenditures due to less upcoding. Further, beyond the financial savings which can be realized through changing individuals' reporting behavior when introducing audits, fines for detected fraudulent behavior under audits help to reduce the losses insurers face due to upcoding. When considering a real-world health setting, it remains unclear whether the detection of dishonest behavior would imply psychological costs, for example derived by concerns for social reputation or self-respect (e.g., Bénabou and Tirole, 2006) and observed lying aversion (e.g., Dufwenberg and Dufwenberg, 2018; Gneezy et al., 2018), which could vary between the lab and the field and might thus lead to different kinds of upcoding behavior. While evaluating costs and benefits is at the discretion of health policy-makers, our findings at least provide some directional guidance for the ongoing debates on the design and implementation of audit policies. Our experimental study suggests interesting paths for future research; for example, in the form of an analysis of reporting behavior under a high detection probability without financial consequences in case of detection, and situations in which upcoding does not yield financial gains. In these ways, the preferences of individuals for dishonesty could be investigated further.

In sum, our results suggest that audits with fines can, on the aggregate, reduce upcoding while not necessarily inducing more honesty. Audits might still decrease dishonesty by pushing dishonest individuals into reporting fraudulently to an extent that is not detectable. This calls for a design of audit policies that makes the detection of dishonest behavior more likely, for example, through audit mechanisms that reduce measurement errors.

Chapter 2

Physicians' incentives, patients' characteristics, and quality of care: A systematic experimental comparison of fee-for-service, capitation, and pay for performance

2.1 Introduction

Paying physicians for performance has become prominent among health policy-makers, for example in the USA (e.g., Rosenthal et al., 2006; Stokes et al., 2018; Song et al., 2019) and in the UK (e.g., Roland, 2004; Doran et al., 2006; Roland and Campbell, 2014). Performance pay (P4P) is usually granted if a quality threshold is reached. Traditional physician payment systems are lump-sum capitation (CAP) or fee-for-service (FFS), in which physicians receive a fee for each service provided, with FFS typically being used in specialty care and CAP being prevalent in primary care (for Germany, see Brosig-Koch et al. 2020). These systems, generally, are not tied to the quality of care provided. FFS incentivizes physicians to overserve patients, whereas CAP embeds an incentive to underserve them. Thus, paying physicians on the basis of direct performance measures has attracted particular attention.

In health care, P4P typically complements either FFS or CAP. From a theoretical point of view, blending P4P with FFS (FFS+P4P) is likely to affect physicians' medical service provision differently compared to P4P blended with capitation (CAP+P4P), due to the different incentives of the baseline payment systems. A systematic comparison of the effectiveness of P4P between CAP and FFS based on comparable designs is lacking. Also, it is not well understood how patients with different severities of illness are affected by incentives of the P4P systems. The heterogeneous impact of payment incentives on different patient types has been indicated in recent empirical (e.g., Clemens and Gottlieb, 2014) and

experimental studies (e.g., Hennig-Schmidt et al., 2011; Brosig-Koch et al., 2017a).

The empirical evidence on whether and, if so, how P4P affects physicians' medical service provision and quality of care, is rather mixed (e.g., Scott et al., 2011; Emmert et al., 2012; Eijkenaar et al., 2013; Milstein and Schreyögg, 2016). Moreover, it has been argued that the *design* of a P4P system is key to effectively changing physician behavior (Epstein, 2012; Maynard, 2012; Kristensen et al., 2016; Anselmi et al., 2020). Potential reasons for the difficulty in establishing a causal link between performance pay and physicians' provision behavior comprise the likely endogeneity of institutions (e.g., Baicker and Goldman, 2011), the biased and incomplete performance measures (e.g., Mullen et al., 2010), measurement errors (e.g., Campbell et al., 2009), the limited availability of data (e.g., Gravelle et al., 2010; Maynard, 2012), and introduction of P4P in parallel to other interventions (e.g., Lindenauer et al., 2007).

Our study contributes to better understanding the effects of different P4P systems on the quantity and quality of care. To this end, we designed a controlled behavioral experiment, in which the physicians' financial incentives in baseline FFS and CAP are mirror images of each other. We complement FFS and CAP with a discrete bonus that is kept constant across both payment systems FFS+P4P and CAP+P4P. The bonus is paid when a quality threshold tied to a patient's optimal health outcome is reached. Meeting the quality threshold still allows for over- and underprovision, as we assume asymmetric information between the physician and the payer. Service provision according to the threshold thus might increase the physician's profit while still not providing the optimal care. This mirror design allows us to systematically compare the two blended payment schemes (FFS+P4P and CAP+P4P) – an analysis that is currently missing in the literature. In addition, we keep the patient population constant. Physicians are confronted with identical patients regarding their severities of illness and their marginal health benefit from each medical service provided. This feature allows us to investigate systematically whether the effects of P4P are specific to patients' illnesses and severities of illness despite the mirror design of the payment systems. Finally, we also consider health policy implications, including cost-benefit analyses, for our experimental design of performance incentives.

Our experimental design is well grounded in theory. Behavioral predictions are derived

from an illustrative model and are tested with physicians in lab-in-the-field conditions and with medical and non-medical students in lab experiments. To establish the causal link between P4P and the quantity and quality of medical service provision, we exogenously vary physicians’ remuneration at a within-subject level from the baseline non-blended payment schemes to the blended performance-pay systems. In a medically framed decision situation, subjects decide on the quantity of medical services for abstracted patients with different severities of illness (mild, intermediate, severe) and marginal health benefits (low, high). Quantity choices determine the physicians’ own profit and the patients’ health benefits measured in monetary terms. Participants are informed that their decisions affect the health of real-world patients, as the money corresponding to the aggregated health benefits is transferred to a charity and is used exclusively for surgery of cataract patients. For an analogous procedure, see, for example, Hennig-Schmidt et al. (2011), Brosig-Koch et al. (2016, 2017a, 2020), and Waibel and Wiesen (2021).

With our parsimonious experiment, we address the following research questions. First, we analyze how the effect of introducing P4P affects medical service provision and the quality of care when complementing FFS. Second, we study whether such an effect is specific to the patients’ characteristics such as the severity of illness and the marginal health benefit. Third, we explore the effect of P4P blended with baseline CAP, and fourth, we investigate the effect differences that are due to the patients’ health characteristics. Finally, in a joint analysis, we examine whether potential differences in subjects’ reactions to FFS and CAP exist and whether the P4P effect varies between FFS+P4P and CAP+P4P, despite the mirror-image design of financial incentives.

Our behavioral results indicate that the introduction of P4P reduces non-optimal service provision, enhances the quality of care, and patients’ health benefit under FFS and under CAP. We find that the effects of P4P are specific to the patients’ severities of illness. Under FFS, the marginal benefit of P4P on medical service provision, the quality of care as well as the patients’ health benefit decreases in patients’ severity of illness. Under CAP, we observe the reverse pattern: the marginal benefit of P4P increases with increasing severity. In other words, our behavioral results indicate that the introduction of pay for performance is most beneficial for mildly ill patients under FFS, whereas it is most beneficial for highly

ill patients under CAP. Patients of intermediate severity of illness are almost equally treated under both performance-pay systems.

While our results suggest that P4P serves as a means to counteract misaligned financial incentives for overprovision under FFS and underprovision under CAP, they also emphasize the importance of its design elements. Utilizing the symmetric design components across baseline payment conditions, we are also able to analyze the cost effectiveness of introducing P4P under both payment conditions and derive health-policy implications. In sum, health policy-makers need to take into account that the effectiveness of P4P is specific to a patient’s severity of illness and the underlying baseline payment condition when designing P4P systems.

This paper proceeds as follows. Section 2.2 lists the streams in the health economics literature to which our paper contributes. Section 2.3 describes our experimental design and behavioral hypotheses. Section 4.5 presents the behavioral results. In Section 2.5, we discuss implications of our results and potential limitations. Section 4.6 concludes.

2.2 Contribution to the literature

We contribute to several streams in the health economics literature. First, we complement the empirical literature that analyzes the effects of P4P on physicians’ treatment decisions. Quite often, P4P programs are evaluated using administrative longitudinal data. Empirical evidence is rather mixed, showing only modest positive effects (if at all) on the quality of medical service provision; for extensive literature reviews, see Scott et al. (2011) for primary care, Jia et al. (2021) for general outpatient care, and Mathes et al. (2019) for inpatient care. Mullen et al. (2010), for example, using longitudinally data from quarterly performance reports, find only little empirical support for a positive effect of introducing P4P on process quality of multi-specialty medical groups in the US. Studies mostly evidence some increase in a few clinical processes; yet, the P4P effects on outcome quality are not clear (e.g., Peckham and Wallace, 2010; Li et al., 2014). While empirical studies typically rely on aggregated data, we add insights on a causal effect of P4P at the individual subject level. The highly controlled environment in our experiment allows us to implement “clean” measures for the

quality of medical service provision of the individual physicians. It also enables us to analyze systematically how variations in patients' health characteristics and payment systems (CAP versus FFS) relate to the effect of P4P.

Second, our study contributes to the scarce experimental literature analyzing performance pay for physicians by means of controlled behavioral experiments. These studies provide first evidence for a positive effect of P4P on physicians' treatment behavior (e.g., Brosig-Koch et al., 2020; Oxholm et al., 2021). Our study differs, however, from this literature by systematically analyzing the effects of FFS *and* CAP as well as of the respective blended P4P systems. The study by Brosig-Koch et al. (2020), with a representative primary-care physician sample, investigates the effect of a threshold-based P4P system with a discrete bonus blended with CAP, analogously to our CAP+P4P condition. A lab experiment with Danish medical students by Oxholm et al. (2021) shows that P4P affects the allocation of medical care across patients with low and high responsiveness to treatment compared to lump-sum CAP payments. Considering FFS and P4P, Keser et al. (2014) report from a laboratory experiment with German medical students that a bonus tied to the share of optimally treated patients leads to some increase in the quality of care. In a lab experiment with US medical students, Cox et al. (2016b) find that utilizing P4P mechanisms incentivizes cost-effective reductions in hospital re-admissions.

Third, we add to the literature on behavioral experiments in health (Galizzi and Wiesen, 2017, 2018), focusing on incentives and physician behavior. In particular, our study complements experiments in a medical framework analyzing the effects of financial incentives on physician behavior, such as FFS or CAP (Hennig-Schmidt et al., 2011; Hennig-Schmidt and Wiesen, 2014; Green, 2014; Brosig-Koch et al., 2016; Lagarde and Blaauw, 2017; Di Guida et al., 2019; Martinsson and Persson, 2019; Reif et al., 2020) and blended payment systems (Brosig-Koch et al., 2017a, 2020). In a broader sense, we also relate to the experimental literature on credence goods markets, which typically apply neutral framings but for which health care characterized by high information asymmetries is a key example (e.g., Dulleck

and Kerschbamer, 2006; Dulleck et al., 2011).¹⁵ Our study adds causal evidence on how, in P4P systems, physician behavior is affected by the baseline payment (FFS versus CAP), how patient characteristics influence treatment decisions, and which design features of a payment system could potentially be implemented to enhance the quality of care for different patient types.

2.3 Experimental design, protocol, and hypotheses

2.3.1 Decision situation

Our experiment employs a medical frame. All subjects decide in the role of physicians on the provision of medical services. We employ a within-subject design to analyze the effect of P4P on physicians’ provision of medical services. To this end, each subject makes his or her decisions under non-blended and blended payments. First, subjects are incentivized either by FFS or by CAP, which serve as baseline payments. Second, we introduce physicians’ P4P in addition to the respective baseline payments (FFS+P4P or CAP+P4P). We randomly assign subjects to one of the two experimental conditions.¹⁶

In all payment systems, physician i decides on the quantity of medical services $q \in [0, 10]$ for nine different patients ($j = 1, \dots, 9$). Patients differ in illnesses $k \in \{A, B, C\}$ and in the severity of illnesses $l \in \{x, y, z\}$. Patients are assumed to be passive and fully insured, accepting each quantity of medical services provided by the physician. This is a common assumption in the theoretical health economics literature (for a comprehensive review, for example, see McGuire, 2000), corresponding to the assumption of information asymmetry

¹⁵Medical services are considered as credence goods due to high informational advantages of physicians towards their patients. This enables physicians to exploit their patients, for example through overtreatment under FFS. In our experimental design, we incorporate the “credence goods” problematic by assuming that our patients are passive and accept each quantity of medical services provided by the physician. Typically applying neutral framings, experiments in the credence goods literature showed that overtreatment can be reduced by costly second opinions (Mimra et al., 2016), competition (Huck et al., 2016), and separating treatment from diagnosis and prescription decisions (Greiner et al., 2017). Recent experiments show that monitoring mechanisms with financial consequences reduce overtreatment and the overcharging of patients (Angerer et al., 2021; Hennig-Schmidt et al., 2019; and Groß et al., 2021). We complement these experiments by investigating whether performance-based financial incentives which implicitly rely on monitoring a physician’s performance are capable of coping with non-optimal medical service provision such as overtreatment under FFS.

¹⁶Notice that the general decision situation of our experiment is similar to Hennig-Schmidt et al. (2011), Hennig-Schmidt and Wiesen (2014), Brosig-Koch et al. (2016, 2017a), and Brosig-Koch et al. (2020). In the latter three studies, incentives under FFS (CAP) are the same as in the present paper.

between expert (physician) and customer (patient) in the credence goods literature (e.g., Dulleck and Kerschbamer, 2006). In our experiment, patients' characteristics are the same in all payment conditions. The patient population for which a physician chooses services thus remains constant.

Physician i 's payment is $R(q) = L + pq + b_l I_{b_l}$, with L being the lump-sum payment, p the fee per service rendered to a patient, and b_l the bonus payment; I_{b_l} denotes an indicator variable which equals 1, if the physician's chosen quantity does not differ by more than one unit from the patient's optimal treatment, and 0 otherwise. In FFS, $L = 0$ and $b_l^{\text{FFS}} = 0$ and in CAP $p = 0$, and $b_l^{\text{CAP}} = 0$.

Physician i 's profit is given as

$$\pi(q) = L + pq + b_l I_{b_l} - c(q), \quad (2.1)$$

with $L, p, b_l \geq 0$, $c'(q) > 0$ and $c''(q) > 0$. In the experiment, $c(q) = q^2/10$ for all payment systems.

When deciding on q , physician i simultaneously determines her own profit $\pi(q)$ and the patient's health benefit $H(q)$ for patient j . Common to all patients' health-benefit functions is a global optimum at q^* on $q \in (0, 10)$. The patient health-benefit function employed in our experiment is

$$H(q) = \begin{cases} H_0 + \theta q & \text{if } q \leq q^* \\ H_1 - \theta q & \text{if } q \geq q^*, \end{cases} \quad (2.2)$$

with $H_0, H_1 \geq 0$ and $\theta > 0$.¹⁷ In particular, for illnesses A and B $\theta = 1$, and for illness C $\theta = 2$. For illnesses A , B , and C , the maximum health benefit is $H_A(q^*) = 7$, $H_B(q^*) = 10$,

¹⁷Note that $H_1 = H_0 + 2\theta q^*$. H_0 and H_1 are allowed to be different, which reflects the patient health benefit parameters in the experiment. For example, for illness A (with $\theta = 1$) and severity x (with $q^* = 3$), $H_0 = 4$ and $H_1 = 10$, as $H_1 = 4 + 2 \cdot 1 \cdot 3 = 10$.

and $H_C(q^*) = 14$, respectively.¹⁸

The patient-optimal quantity q^* depends on a patient’s severity of illness l . For mild (x), intermediate (y), and high (z) severe illnesses, the patient-optimal quantities are $q_x^* = 3$, $q_y^* = 5$, and $q_z^* = 7$, respectively. Varying patients’ characteristics in our experiment are motivated by the recent theoretical literature (see, e.g., Allard et al., 2011), which assumes that patients’ characteristics affect the physicians’ behavior. Figure B.1.2 in Appendix B.1.3 illustrates the patient health benefits in our experiment, which are varied systematically for the patients’ illness k and severity of illness l . The differences in optimal quantities and marginal health benefits by patients’ characteristics are motivated by recent claims for more value-based health care which focuses on patients’ needs. In the experiment, the patient-optimal quantity q^* for all patients is common knowledge, so are all parameters of the experiment. Thus, when making their quantity choices, physicians are aware of cost, payment, profit, and the patient’s health benefit for each quantity; for an illustration of the decision situation, see the instructions in Appendix B.1.4.¹⁹

Our experimental design enables us to investigate how different payment schemes and performance-based payment components, which are linked to the generated health benefit (health outcome), affect treatment decisions. We are able, first, to analyze the quantity of medical services and, second, to introduce a “clean” quality measure related to the patient-optimal treatment. Moreover, the symmetric design of patient health benefits implies that the marginal effects (i.e., the absolute value of $H'(q)$) of over- and underprovision of medical services are equivalent. This parsimonious design with mirror-image incentives allows for a systematic comparison of incentives from P4P on the quantity and quality of

¹⁸Patients’ health benefits are measured in monetary terms. The accumulated benefits are then transferred to a charity that supports surgical treatment of real cataract patients. Note that this “mechanism” implies that a monetary amount deriving from subjects’ decisions in the lab is applied to the treatment of real patients, which makes it different from the kinds of donations analyzed in the charitable-giving literature; see, for example, Andreoni (1989) or DellaVigna et al. (2012). This procedure, which was introduced by Hennig-Schmidt et al. (2011), has been used in several experiments in health economics, as it embeds an incentive for subjects in the experiment that relates to real patients’ health in the real world. Equivalent mechanisms have been employed in recent behavioral experiments in the field of health care which have analyzed physician behavior (Hennig-Schmidt and Wiesen, 2014; Godager et al., 2016; Brosig-Koch et al., 2016, 2017a, 2020; Byambadalai et al., 2019; Di Guida et al., 2019; Martinsson and Persson, 2019; Huesmann et al., 2020; Waibel and Wiesen, 2021; Wang et al., 2020). In Kesternich et al. (2015) and Lagarde and Blaauw (2017), subjects could choose from several (medical) charities to which a donation could be transferred.

¹⁹This allows a clean analysis of the extent to which patient-regarding concerns guide physicians’ medical service provision, while excluding potential additional influences like risk preferences.

care and a systematic cost-benefits analysis.

2.3.2 Payment systems

Recall that each subject decides in the role of a physician on the provision of medical services under non-blended and blended payment systems. Table 2.1 provides an overview of payment systems employed in our experiment. In part *I* of the experiment, subjects decide either under FFS or CAP. Subjects paid by FFS (CAP) in part *I* decide under the associated P4P system (FFS+P4P or CAP+P4P) in part *II*. The profit functions of FFS and FFS+P4P systems mirror those of the respective CAP and CAP+P4P systems. While varying the components of the payment systems, we keep maximum profit levels and marginal profits constant. The profit parameters are illustrated in Figure B.1.3, and the complete set of parameter values is shown in Table B.1.2 in Appendix B.1.3.

In FFS, physicians are paid a fee of $p = 2$ per service. Accordingly, profit is $\pi(q) = 2q - c(q)$. In CAP, physicians receive a lump-sum payment of $L = 10$ per patient, independently of the quantity of medical services. Physicians' profit per patient is thus $\pi(q) = 10 - c(q)$ with the maximum attainable profit being 10 in both payment systems FFS and CAP. The profit-maximizing quantity of medical services for each of the nine patients is $\hat{q}_j^{\text{FFS}} = 10$ and $\hat{q}_j^{\text{CAP}} = 0$ in FFS and CAP, respectively. This reflects the prevalent financial incentives for overprovision under FFS and underprovision under CAP.

Table 2.1: Experimental parameters

First part of the experiment (Non-blended payment systems)				Second part of the experiment (Blended payment systems)						Subjects (physicians, medical students, non-medical students)
Payment	L	p	R	Payment	Severity l	L	p	b_l	R	
FFS	–	2	$2q$	FFS+P4P	x	–	2	5.6	$2q + 5.6$	52 (10, 22, 20)
					y	–	2	3.6	$2q + 3.6$	
					z	–	2	2.4	$2q + 2.4$	
CAP	10	–	10	CAP+P4P	x	10	–	2.4	$10 + 2.4$	55 (10, 22, 23)
					y	10	–	3.6	$10 + 3.6$	
					z	10	–	5.6	$10 + 5.6$	

Notes. This table shows the parameters and the number of participants in each experimental part. Note that the performance pay b_l is only granted to subjects if their quantity choice fulfills the quality requirement $|q - q^*| \leq 1$; otherwise the performance pay equals zero. Data for the non-blended payment systems correspond to a part of the data analyzed in Brosig-Koch et al. (2016).

Our performance measure is linked to a patient’s health outcome – namely, the optimal patient health benefit. P4P is granted if the quantity chosen by a physician does not deviate by more than one unit from the patient-optimal quantity q^* ; i.e., whenever $|q - q^*| \leq 1$. We thereby assume that the quality is not fully contractible due to information asymmetry. P4P systems, thus, mitigate inherent incentives to provide too many services under FFS and too few under CAP, respectively. In our experiment, we determine the profit-maximizing quantities under P4P such that they are “closer” to the patient-optimal quantities than in non-blended FFS or CAP, but do not coincide with them. Since the design of performance-based bonus payments incentivizes the smallest possible deviation from q^* instead of q^* itself, we are also able to differentiate between profit maximization and optimal patient care in our P4P conditions.

We set bonus rates such that incentives are comparable across payment systems. For severities x , y , and z , $b_x^{\text{FFS}} = 5.6$, $b_y^{\text{FFS}} = 3.6$, $b_z^{\text{FFS}} = 2.4$ in FFS+P4P, and $b_x^{\text{CAP}} = 2.4$, $b_y^{\text{CAP}} = 3.6$, $b_z^{\text{CAP}} = 5.6$ in CAP+P4P, respectively. The bonus implies an increase in the maximum attainable profit $\pi(\hat{q}_j)$ by 20 percent. For each severity, choosing \hat{q}_j equal to 4, 6, or 8 (2, 4, or 6) in FFS+P4P (CAP+P4P) thus yields a profit of 12 for the physician.

2.3.3 Experimental protocol

Overall, 107 subjects participated in our experiment. Among these were 44 medical and 43 non-medical students who took part in the lab experiments and 20 physicians who took part in artefactual field experiments. Each subject was randomly assigned to only one of the two baseline payment systems. In particular, 55 subjects took part in CAP/CAP+P4P and 52 in FFS/FFS+P4P; with 22 medical students and 10 physicians under each payment system; see Table 2.1.

The computerized experiment was programmed with z-Tree (Fischbacher, 2007). Physicians and students were presented with identical computer screens, instructions, and comprehension questions. The only differences were a higher exchange factor from the experimental currency to Euro for the physicians’ payoffs compared to the students’ payoffs and minor

deviations in the experimental procedure.²⁰ The artefactual field experiments were conducted in 2012 and 2013 using the mobile lab of the Essen Laboratory for Experimental Economics (elfe) at the Academy for Training and Education of Physicians (Akademie für Ärztliche Fort- und Weiterbildung) in Bad Nauheim, Germany. At the Academy, German physicians contracting with the statutory health insurers take mandatory annual education and training courses. The physicians were recruited by announcements in their courses. They voluntarily participated before or after their courses. The lab experiments were conducted between 2011 and 2013 at elfe at the University of Duisburg-Essen.²¹ Student subjects were recruited via the online recruiting system ORSEE (Greiner, 2015).

The experimental procedure was as follows: Upon arrival, subjects were randomly assigned to workstations separated by panels to ensure that decisions could be made in full anonymity. They then were given ample time to read the instructions for part *I*. Subjects were informed that the experiment consisted of two parts, but received detailed instructions for part *II* only after having finished part *I* of the experiment. To check for the subjects' understanding of the decision task, they had to answer a set of control questions. The experiment did not start unless all subjects had answered the control questions correctly. Instructions are to be found in Appendix B.1.4. In each of the two parts of the experiment, subjects then subsequently decided on the quantity of medical services for each of the nine patients, i.e., for each possible combination of illnesses and severities. The order of patients was randomly determined and kept constant for all subjects and all conditions: $Bx, Cx, Az, By, Bz, Ay, Cz, Ax, Cy$.

Before making their decision for a specific patient, subjects were informed about their payment, their cost and profit, as well as about the patient benefit for each quantity from 0 to 10. All monetary amounts are given in Taler, our experimental currency. The exchange rate is 1 Taler = EUR 0.80 in the lab experiment and 1 Taler = EUR 3.40 in the artefactual field experiment. Compared to the lab, the payment in the field experiment was increased

²⁰Before the experiments, physicians were briefly introduced to the experimental economics method, the universities involved in running the experiment, and the funding institution of our research project (DFG, German Research Foundation). After the experiment, physicians were debriefed and informed about results of previous health-related economic experiments.

²¹For a picture of the setup of the mobile lab in Bad Nauheim and the typical setup of the computer laboratory at elfe, see B.1.1 in Appendix B.1.1.

by a factor of 4.25 to provide adequate incentives for the physicians.²² The procedure was exactly the same in part *II* of the experiment. After finishing part *II*, we asked the subjects to complete a questionnaire on social demographics (e.g., age and gender) and on personality traits elicited by a ten-item personality inventory, which comprises five personality dimensions: extraversion, agreeableness, conscientiousness, neuroticism, and openness (Rammstedt and John, 2007). An overview on summary statistics on social demographics and personality traits can be found in Table B.1.1 in Appendix B.1.2.

At the end of the experiment, when all subjects had made their decisions, we randomly determined one decision in each part of the experiment to be relevant for a subject’s actual payoff and the patient benefit. This was done to rule out income effects. Subjects were paid in private according to these two randomly determined decisions.

To verify that the money corresponding to the sum of patient benefits in a session was actually transferred to the charity, we applied a procedure similar to Hennig-Schmidt et al. (2011) and Brosig-Koch et al. (2016, 2017a). One of the participants was randomly chosen to be the monitor. After the experiment, the monitor verified that a payment order on the aggregated benefit in the respective session was written to the financial department of the University of Duisburg-Essen to transfer the money to the Christoffel Blindenmission, which used the monetary transfers exclusively to support surgical treatments of cataract patients in a hospital in Masvingo (Zimbabwe) staffed by ophthalmologists from the charity.²³ The order was sealed in an envelope and the monitor and the experimenter then walked together to the nearest mailbox and deposited the envelope. The monitor was paid an additional EUR 5.

Laboratory sessions lasted for about 60 minutes. Subjects earned, on average, EUR 16.37. The average benefit per patient was EUR 13.25. In total, EUR 1,152.80 were transferred to the Christoffel Blindenmission. The average cost for a cataract operation amounts, according

²²The amount physicians could earn in the experiment was set such that it reflects the average net hourly wage of a physician in Germany, bearing in mind potential differences, for example across the physicians’ specialization and seniority. We set this factor after consultation with Dr. Harald Herholz of the Association of Statutory Health Insurance Physicians in Hesse (Germany), who has been involved in budget negotiations for physicians’ remuneration.

²³Notice that we did not inform the subjects that the money was assigned to a developing country. We wanted to avoid motives like compassion for people in developing countries that are independent of being in need of ophthalmic surgery. Feedback from the subjects in a pre-experimental pilot session in Hennig-Schmidt et al. (2011) actually raised this issue.

to the Christoffel Blindenmission, to about EUR 30. Thus, our experiment allowed 38 cataract patients to be treated. The sessions of the artefactual field experiment lasted for about 50 minutes. Physicians earned, on average, EUR 62.73. The average benefit per patient was EUR 67.83. In total, EUR 1,356.60 were transferred to the Christoffel Blindenmission, allowing the treatment of 45 cataract patients.

2.3.4 Behavioral hypotheses

To organize our thoughts, we now describe the physicians' behavior more formally and derive behavioral predictions for our experiment. To this end, we follow the intuition of an illustrative model by Brosig-Koch et al. (2017a). The formal model is relegated to Appendix B.2. We assume that a physician derives utility from her own profit and from a patient's health benefit. The weight the physician attaches to the patient's health benefit is interpreted as a measure for physician altruism. The assumption of physicians being altruistic has become common in the health economics literature, since Arrow (1963) coined the importance of physicians' patient-regarding motivation in the delivery of medical services.²⁴

First, we consider a physician's behavior under the baseline payment systems FFS and CAP. For the profit and patient benefit parameters in our experiment and the given altruism of a physician, we conjecture that FFS induces overprovision of medical services, which decreases in the severity of a patient's illness and in the patient's marginal health benefit.

On the contrary, we expect that CAP induces underprovision of medical services, which increases in the severity of illness, and decreases in the marginal health benefit. Ample evidence for these conjectures on effects of FFS and CAP exists from related experiments (e.g., Hennig-Schmidt et al., 2011; Brosig-Koch et al., 2016, 2017a; Martinsson and Persson, 2019; Brosig-Koch et al., 2020). With a higher severity of illness, more medical services are provided (Hennig-Schmidt et al., 2011; Brosig-Koch et al., 2016, 2017a; Brosig-Koch et al., 2020). These behavioral effects related to the severity of illness are particularly relevant in our experiment, as the levels of P4P are tied to the patients' severity of illness; for an

²⁴In addition to the importance for designing optimal payment schemes (e.g., Ellis and McGuire, 1986, 1990; Ma, 1994; Chalkley and Malcomson, 1998; Jack, 2005; Choné and Ma, 2011; Olivella and Siciliani, 2017), a physician's altruism is essential, for example in analyzing referral decisions (Allard et al., 2011; Waibel and Wiesen, 2021), responses to transparency (Kolstad, 2013), prescription of generic drugs (e.g., Hellerstein, 1998; Crea et al., 2019), and the delegation of treatment decisions (Liu and Ma, 2013).

illustration, see Figure B.1.3 in Appendix B.1.3.²⁵

Our main focus is on the effect of P4P. When introducing P4P the bonus b_l is granted if and only if a physician’s treatment decision meets the quality threshold $|q - q^*| \leq 1$ for a patient with a severity of illness l . Recall that we thus assume that the quality is not fully contractible due to information asymmetry between physician and payer. By linking performance pay to the optimal health benefit, the interests of the physician and the patient become (more) aligned. While P4P incentivizes less altruistic physicians to provide medical services ‘close’ to the patient-optimal quantity, baseline financial incentives for underprovision under CAP and overprovision under FFS are still inherent (albeit to a substantially lower extent). Hence, we conjecture that P4P reduces overprovision of medical services in FFS and underprovision in CAP. For the proof, see Appendix B.2.

Intuitively, whether a physician under baseline FFS and CAP chooses a quantity of medical services corresponding to the quality threshold depends on physician i ’s degree of altruism towards the patient, which counterbalances the incentive effects of FFS and CAP. Given a physician’s altruism $\alpha_i \in [0, 1)$, we therefore expect P4P to reduce non-optimal service provision under FFS+P4P and CAP+P4P. Previous experimental evidence (Hennig-Schmidt et al., 2011; Brosig-Koch et al., 2016, 2017a; Brosig-Koch et al., 2020) has shown that, in the basic payment schemes FFS and CAP, non-optimal service provision is highest for patients where \hat{q} and q^* are most misaligned. This is the case for mild-severity patients under FFS and high-severity patients under CAP. We therefore expect the effect sizes of P4P to vary with patients’ severities of illness. In sum, we state the following hypotheses about physicians’ medical service provision and the quality of care.

Hypothesis 1 *P4P blended with FFS.*

A threshold-based performance pay system with a discrete bonus reduces the overprovision of medical services under fee-for-service and increases the quality of care.

²⁵Performance pay for mild-severity patients, b_x^{FFS} , is highest with $b_x^{\text{FFS}} > b_y^{\text{FFS}} > b_z^{\text{FFS}}$, as the “risk” of a mild-severity patient being overserved and therefore suffering disutility is highest under FFS. In CAP, the incentive to undertreat patients increases in the severity of illness; therefore, $b_z^{\text{CAP}} > b_y^{\text{CAP}} > b_x^{\text{CAP}}$ (see Table 2.1).

Hypothesis 2 *FFS+P4P and patient's health characteristics.*

Under performance pay and fee-for-service, the effect of performance pay on medical service provision and the quality of care decreases in the patient's severity of illness and the patient's marginal health benefit.

Hypothesis 3 *P4P blended with CAP.*

Introducing performance pay reduces the underprovision of medical services under capitation and enhances the quality of care.

Hypothesis 4 *CAP+P4P and patient's health characteristics.*

Under performance pay and capitation, the effect of performance pay increases in the patient's severity of illness and the patient's marginal health benefit.

Following directly from Hypotheses 2 and 4, we state:

Hypothesis 5 *Comparison of FFS+P4P and CAP+P4P.*

FFS+P4P leads to a larger improvement in the quality of care for mildly ill patients compared to CAP+P4P. For severely ill patients, the increase in quality of care is larger for CAP+P4P, while for intermediately ill patients, the quality of care does not differ between the two pay for performance systems.

2.4 Behavioral results

In this section, we first provide an introductory, mostly descriptive analysis of medical service provision and the quality of care under the baseline payment systems, FFS and CAP, and the blended performance pay systems, FFS+P4P and CAP+P4P (Section 2.4.1). Second, we test our Hypotheses 1 and 2 on the effects of performance pay when blended with FFS (Section 2.4.2) and, third, analogously, for a blended CAP and performance pay, we test Hypotheses 3 and 4 (Section 2.4.3). Finally, we compare the effects of blended payment systems FFS+P4P and CAP+P4P according to Hypotheses 5 (Section 2.4.4).

In our analyses, we consider the quantity of medical services q and capture the quality of care considering two quality measures. First, our choice-based measure is the absolute deviation from the patient-optimal quantity $\rho = |q - q^*|$. Second, our outcome-based measure

is the proportional health benefit \hat{H} . It comprises the patient's health benefit realized by the physician's actual quantity choice (H_{kl}) as a proportion of the highest achievable health benefit (H_l^*). To facilitate the comparisons across patients (kl) who also vary in terms of their minimal health benefit H^{\min} , we normalize our measure accounting for the lower bounds of achievable health benefit, and define $\hat{H}_{kl} = \frac{H_{kl}^{\min} - H_{kl}}{H_{kl}^{\min} - H_l^*}$. When physician i provides the patient-optimal quantity q^* , the proportional health benefit is highest and $\hat{H}_{kl} = 1$. \hat{H}_{kl} implies a measurable health outcome which allows us to compare actual with optimal quality of care across the different payment systems.

2.4.1 Introductory analyses

Figure 2.1 illustrates the average quantity under the four payment systems for each of the nine patients who differ by illness k and severity of illnesses l .²⁶ We find that subjects provide significantly more services under FFS (Mean 6.69, s.d. 2.07) than under CAP (Mean 3.32, s.d. 2.13), aggregated over all patients ($p < 0.001$, two-sided Mann-Whitney U-test, MWU in the following), see Panel A of Table B.3.1 in Appendix B.3. This finding is in line with earlier experimental studies (recall Section 2.3.4). In FFS+P4P, the quantities of medical services decrease by about 16.4 percentage points (Mean 5.59, s.d. 1.71), and in CAP+P4P, they increase by about 32.5 percentage points (Mean 4.40, s.d. 1.66).

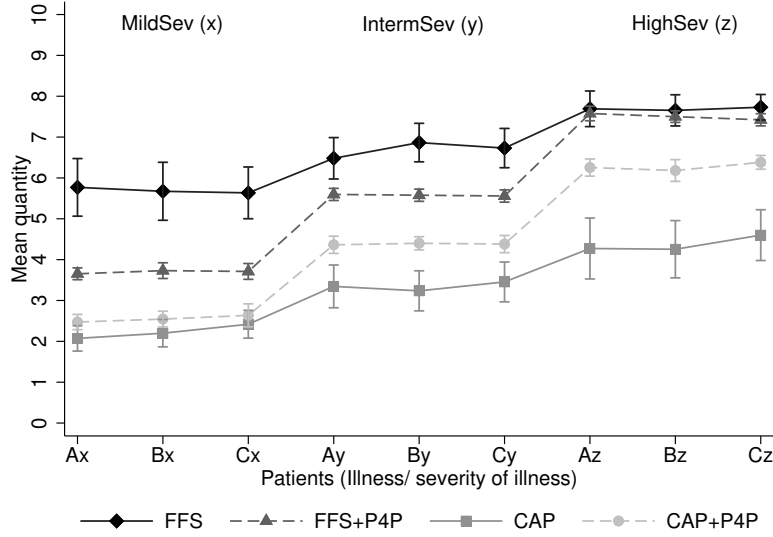
Under FFS, the absolute deviation from the patient-optimal quantity $\rho = 1.82$ (s.d. 1.95), aggregated over all patients. Introducing P4P reduces the average non-optimal service provision ρ to 0.63 (s.d. 0.55), which is a reduction by 65.4 percentage points. Under CAP, $\rho = 1.77$ (s.d. 2.01), while under CAP+P4P ρ declines to 0.65 (s.d. 0.75), a decrease by 63.3 percentage points. See Panel B of Table B.3.1 for detailed descriptive statistics on our choice-based measure ρ .

For the proportional health benefit \hat{H} , we find that, on average, around 71% of the maximum health benefit is realized in the two basic payment systems and around 90% in the two blended payment systems. The effect of P4P thus corresponds to an overall increase in the proportional health benefit by 19 percentage points under CAP+P4P and FFS+P4P.

²⁶See the distributions of medical services differentiated by payment schemes and patients in Figure B.3.1, Appendix B.3.

Detailed descriptive statistics on \hat{H} are provided in Panel C of Table B.3.1 in Appendix B.3. On the aggregate, the introduction of P4P leads to a significant increase in the quantity, and in both the choice-based and the outcome-based quality measures ($p < 0.001$, Wilcoxon signed-rank test, two-sided).

Figure 2.1: Mean quantity by patients' health characteristics



Notes. This figure shows the mean quantity with 95% confidence interval under the four payment systems for each of the nine patients kl . Patients vary by their illness $k = A, B, C$ and severity of illnesses l with mild (x), intermediate (y), and high (z) severe illnesses.

For the proportional health benefit \hat{H} , we find that, on average, around 71% of the maximum health benefit is realized in the two basic payment systems and around 90% in the two blended payment systems. The effect of P4P thus corresponds to an overall increase in the proportional health benefit by 19 percentage points under CAP+P4P and FFS+P4P. Detailed descriptive statistics on \hat{H} are provided in Panel C of Table B.3.1 in Appendix B.3. On the aggregate, the introduction of P4P leads to a significant increase in the quantity, and in both the choice-based and the outcome-based quality measures ($p < 0.001$, Wilcoxon signed-rank test, two-sided).

We further observe that the patients' severity of illness substantially affects the subjects' behavior in all payment systems; see Table B.3.1 in Appendix B.3. Overprovision of medical services is highest for mildly ill patients in both FFS conditions, and underprovision is

highest for severely ill patients in both CAP conditions. The behavioral effect is rather less pronounced for the marginal health benefit. For a detailed overview on descriptive statistics and non-parametric tests for all payment systems and the patients' characteristics, see Table B.3.2 in Appendix B.3.

We now briefly analyze whether responses to the baseline payment systems FFS and CAP are different. To this end, we run a regression on the quantity and quality of care at a between-subject level, see Table B.3.3 in Appendix B.3. In line with the inherent incentives, we find that the treatment quantity is, on average, 3.44 medical services lower under CAP than under FFS. Due to the opposing inherent incentives, effects on the quality of care are more meaningful when comparing both payment systems. On average, neither non-optimal care nor proportional health benefit differs significantly between both baseline payment systems; see Panel B and C of Table B.3.3 in Appendix B.3. Our regression results indicate that the incentives to underprovide in CAP are equally strong as the incentives to overprovide in FFS. For a patient's severity of illness, we find no significant differences in non-optimal care, except in the proportional patient's health benefit, which is on average about 10.0 percentage points lower for an intermediately ill than for a mildly ill patient. Patients with a high marginal health benefit receive on average a significantly higher quality of care than patients with a low marginal health benefit.

2.4.2 The effect of blending fee-for-service with performance pay

We next analyze how introducing P4P+FFS affects the quantity and quality of medical service provision on the individual level. These analyses are particularly important, as a detailed experimental investigation of these effects was lacking up to now.

To estimate the P4P effect, we use OLS regressions for the independent variables q_{ij} (quantity chosen), and $\rho_{ij} = |q_{ij} - q_j^*|$ and a fractional probit response model for the proportional health benefit \hat{H}_{ij} , scaled between 0 and 1. Our base econometric specification is as follows:

$$y_{ij} = \alpha + \beta_1 \text{P4P} + \beta_2 \text{INTERMSEV} + \beta_3 \text{HIGHSEV} + \beta_4 \text{HIGHMHB} + \beta_5 \mathbf{X}_i + \epsilon_{ij}. \quad (2.3)$$

INTERMSEV and HIGHSEV are dummy variables for intermediate and high severities of illness, respectively. HIGHMHB is a dummy for the marginal health benefit being 1 if $\theta = 2$ (high), and 0 otherwise ($\theta = 1$, low). P4P is a dummy variable indicating the introduction of P4P. X_i is a vector of subject i 's characteristics comprising gender, personality traits and medical background (non-medical students, medical students or physicians). We account for potentially confounding effects by medical background as previous experimental evidence (e.g., Brosig-Koch et al., 2016) indicates that behavioral responses to financial incentives might differ (merely in size not qualitatively) by subject pool. Our estimated effects of P4P remain stable when we control for subjects' medical background and other characteristics.²⁷

According to Hypothesis 1, we expect that P4P reduces the quantity of medical services, induces less overprovision, and induces a higher proportional health benefit. Models (1), (4), and (7) in Table 2.2 show that Hypothesis 1 is supported. Introducing FFS-based P4P leads to a highly significant reduction in treatment quantity by, on average, 1.10 medical services. Non-optimal care also declines highly significantly, by 1.20 medical services on average. The proportional patient's health benefit rises by about 18.9 percentage points when P4P is introduced. We summarize the regression results as follows:

Result 1 (P4P blended with FFS) *Complementing fee-for-service with a threshold-based performance-pay system leads to a decrease in overprovision of medical services, which corresponds to an increase in the quality of medical care and in the proportional health benefit.*

Hypothesis 2 considers the effect the severities of illness and the marginal health benefit have on physicians' responses to P4P. Before testing the hypothesis, we analyze the respective impacts on the physicians' treatment decisions under FFS payment conditions in general. Compared to mildly ill patients, treatment quantities increase significantly for intermediately and severely ill patients by, on average, 1.44 and 2.90 medical services, respectively; see Model (1) of Table 2.2. Considering treatment quality, non-optimal care significantly decreases with increasing severity (Model (4)). The proportional health-benefit increase for severely ill patients is significantly higher than for mildly ill patients (by 13.3 percentage points)

²⁷For a comparisons of regression estimates without individual controls, see Tables B.3.4 and B.3.5 in Appendix B.3. For regression estimates with the full list of individual controls, see Tables B.3.6 and B.3.7 in Appendix B.3.

Table 2.2: Regression models on the effect on quantity and quality under FFS conditions

Method: Model:	A. Quantity of medical services q			B. Absolute deviation from optimal care ρ			C. Proportional health benefit \hat{H}		
	OLS (1)	OLS (2)	OLS (3)	OLS (4)	OLS (5)	OLS (6)	Frac. Probit (7)	Frac. Probit (8)	Frac. Probit (9)
P4P	-1.100*** (0.185)			-1.199*** (0.172)			0.189*** (0.025)		
INTERMSEV	1.439*** (0.086)	1.000*** (0.147)	1.439*** (0.086)	-0.529*** (0.086)	-0.936*** (0.149)	-0.529*** (0.086)	0.004 (0.010)	0.019 (0.012)	0.004 (0.010)
HIGHSEV	2.901*** (0.129)	2.000*** (0.230)	2.901*** (0.129)	-0.997*** (0.141)	-1.782*** (0.256)	0.997*** (0.141)	0.133*** (0.016)	0.184*** (0.023)	0.133*** (0.016)
HIGHMHB	-0.016 (0.053)	-0.016 (0.053)	0.010 (0.089)	-0.054 (0.051)	-0.054 (0.051)	-0.074 (0.087)	0.009 (0.008)	0.009 (0.008)	0.008 (0.011)
P4P \times MILDSEV		-1.994*** (0.277)			-1.994*** (0.277)		0.187*** (0.021)		
P4P \times INTERMSEV		-1.115*** (0.192)			-1.179*** (0.183)		0.162*** (0.021)		
P4P \times HIGHSEV		-0.192 (0.132)			-0.423*** (0.111)		0.079*** (0.017)		
P4P \times LOWMHB			-1.083*** (0.195)			-1.212*** (0.179)		0.171*** (0.022)	
P4P \times HIGHMHB			-1.135*** (0.177)			-1.173*** (0.172)		0.153*** (0.018)	
Constant	5.623*** (0.315)	6.070*** (0.350)	5.615*** (0.318)	2.621*** (0.315)	3.019*** (0.354)	2.627*** (0.317)			
Wald test (p -value)									
H_0 : P4P \times MILDSEV = P4P \times INTERMSEV		<0.001			<0.001			0.010	
H_0 : P4P \times MILDSEV = P4P \times HIGHSEV		<0.001			<0.001			<0.001	
H_0 : P4P \times INTERMSEV = P4P \times HIGHSEV		<0.001			<0.001			<0.001	
H_0 : P4P \times LOWMHB = P4P \times HIGHMHB			0.556			0.658			0.062
Observations	936	936	936	936	936	936	936	936	936
Subjects	52	52	52	52	52	52	52	52	52
(Pseudo) R^2	0.563	0.599	0.563	0.336	0.379	0.336	0.150	0.157	0.150

Notes. This table shows parameter estimates from OLS regressions (Panel A and B) and average marginal effects from fractional probit response regressions (Panel C). Robust standard errors clustered for subjects are shown in parentheses. P4P is a dummy variable indicating the introduction of P4P. INTERMSEV and HIGHSEV are dummy variables for intermediate and high severities of illness. HIGHMHB is a dummy for the marginal health benefit being 1 if $\theta = 2$ (high), and 0 otherwise ($\theta = 1$, low). All models control for individual characteristics which comprise gender, medical background (non-medical student, medical student, physician), and personality traits; for the respective estimates, see Table B.3.6 in Appendix B.3. * $p < 0.10$, ** $p < 0.05$, and, *** $p < 0.01$.

but does not significantly differ between intermediately and mildly ill patients (Model (7)). These findings are in line with results reported by, e.g., Brosig-Koch et al. (2017a) and Martinsson and Persson (2019).

Hypothesis 2 states that the effect of P4P on medical service provision and on the quality of care decreases in the severity of illness and in the marginal health benefit. To estimate the moderating effects that patients' severities of illness have on responses to P4P, we consider the following model:

$$\begin{aligned}
y_{ij} = & \alpha + \beta_1 \text{INTERMSEV} + \beta_2 \text{HIGHSEV} + \beta_3 \text{HIGHMHB} + \beta_4 \text{P4P} \mathbf{x} \text{MILDSEV} \\
& + \beta_5 \text{P4P} \mathbf{x} \text{INTERMSEV} + \beta_6 \text{P4P} \mathbf{x} \text{HIGHSEV} + \beta_7 \mathbf{X}_i + \epsilon_{ij}.
\end{aligned} \tag{2.4}$$

Following an econometric approach by Clark and Huckman (2012), we include the terms $\beta_4 \text{P4P} \mathbf{x} \text{MILDSEV}$, $\beta_5 \text{P4P} \mathbf{x} \text{INTERMSEV}$, and $\beta_6 \text{P4P} \mathbf{x} \text{HIGHSEV}$, which interact P4P with each severity level of illness to determine the extent to which the effect (marginal benefit) of P4P depends on the patient's severity of illness. By construction, the estimates of β_4 , β_5 , and β_6 represent the total effect of P4P for patients with either a mild severity of illness, an intermediate severity of illness or a high severity of illness, respectively.

Regression results provide support for Hypothesis 2 (see Models (2), (5), and (8), as well as Wald test results). First, P4P positively affects the quantity and quality of care, as all coefficients on the effects are significantly different from zero, except the effect of P4P on the quantity of medical services for severely ill patients (see Model (2)). Second, we find the anticipated relation between severity of illness and P4P such that coefficients are significantly higher for less severely ill patients.

For a patient's level of marginal health benefit, we neither find a significant effect on the quantity of medical service provision nor on the quality of care, see Models (1), (4), and (7). To estimate whether the positive effect of P4P differs between patients with high and low marginal benefits, we consider a model similar to Equation (2.4), in which we interact P4P with each marginal health-benefit level. When comparing the effect of P4P for patients with a low marginal health benefit ($\text{P4P} \mathbf{x} \text{LowMHB}$) to the effect for patients with a high marginal health benefit ($\text{P4P} \mathbf{x} \text{HighMHB}$), we observe no significant differences; see Models

(3), (6), and (9) of Table 2.2 and the Wald tests. We summarize our findings as follows:

Result 2 (FFS+P4P and patients' health characteristics) *While introducing performance pay improves the quality of medical service provision across all severity types, the effect of performance pay significantly decreases with increasing severity of illness. For patients' marginal health benefit, the effect of performance pay is less systematic.*

2.4.3 The effect of blending capitation with performance pay

We now analyze the effects of introducing P4P to CAP on the quantity and quality of care. An earlier study by Brosig-Koch et al. (2020) used the same design to investigate the P4P effect with general practitioners and medical students when CAP is the baseline payment. We repeat the analyses with our data according to our econometric specifications in Equations (3.1) and (2.4). We thus provide the basis for jointly comparing the payment systems FFS, CAP, FFS+P4P, and CAP+P4P in Section 2.4.4.

According to Hypothesis 3, we expect that introducing P4P to CAP increases the treatment quantity, reduces the underprovision of medical services, and enhances the quality of care. Our data support Hypothesis 3. Models (1), (4), and (7) of Table 2.3 show that CAP+P4P leads to a highly significant increase in the treatment quantity by on average 1.09 services, a reduction of non-optimal care by on average 1.12 medical services, and an increase in the proportional health benefit by about 17.5 percentage points.²⁸ In sum, we state:

Result 3 (P4P blended with CAP) *Complementing capitation with performance pay leads to an increase in medical services, a decrease in underprovision, and a rise in patients' proportional health benefit.*

To analyze the effects severities of illness and the marginal health benefit have on the physicians' responses to CAP+P4P (Hypothesis 4), we again first study the impact the severities of illness have on the physicians' treatment decisions, as indicated in Equation (3.1). We find that treatment quantities for intermediately and severely ill patients are significantly higher than for mildly ill patients by, on average, 1.47 and 2.93 medical services,

²⁸Regression estimates for models without individual controls yield very similar results; see Table B.3.5 in Appendix B.3.

Table 2.3: Regression models on the effect on quantity and quality under CAP conditions

Method: Model:	A. Quantity of medical services q			B. Absolute deviation from optimal care ρ			C. Proportional health benefit \hat{H}		
	OLS (1)	OLS (2)	OLS (3)	OLS (4)	OLS (5)	OLS (6)	Frac. Probit (7)	Frac. Probit (8)	Frac. Probit (9)
P4P	1.085*** (0.189)			-1.117*** (0.180)			0.175*** (0.026)		
INTERMSEV	1.473*** (0.100)	1.115*** (0.139)	1.473*** (0.100)	0.436*** (0.074)	0.848*** (0.138)	0.436*** (0.074)	-0.143*** (0.016)	-0.201*** (0.024)	-0.143*** (0.016)
HIGHSEV	2.933*** (0.151)	2.145*** (0.245)	2.933*** (0.151)	0.958*** (0.134)	1.758*** (0.250)	0.958*** (0.134)	-0.149*** (0.019)	-0.227*** (0.028)	-0.149*** (0.019)
HIGHMHB	0.179*** (0.048)	0.179*** (0.048)	0.261*** (0.069)	-0.115** (0.044)	0.115** (0.044)	-0.194** (0.073)	0.017** (0.007)	0.017** (0.007)	0.024*** (0.009)
P4P \times MILDSEV		0.321** (0.140)			-0.309*** (0.108)			0.055*** (0.017)	
P4P \times INTERMSEV		1.036*** (0.201)			-1.133*** (0.189)			0.157*** (0.021)	
P4P \times HIGHSEV		1.897*** (0.296)			-1.909*** (0.292)			0.180*** (0.023)	
P4P \times LOWMHB			1.139*** (0.195)			-1.170*** (0.188)			0.165*** (0.024)
P4P \times HIGHMHB			0.976*** (0.193)			-1.012*** (0.173)			0.135*** (0.018)
Constant	1.725*** (0.250)	2.107*** (0.231)	1.698*** (0.254)	1.440*** (0.242)	1.036*** (0.227)	1.466*** (0.247)			
Wald test (p -value)									
H_0 : P4P \times MILDSEV = P4P \times INTERMSEV		<0.001			<0.001			<0.001	
H_0 : P4P \times MILDSEV = P4P \times HIGHSEV		<0.001			<0.001			<0.001	
H_0 : P4P \times INTERMSEV = P4P \times HIGHSEV		<0.001			<0.001			0.015	
H_0 : P4P \times LOWMHB = P4P \times HIGHMHB			0.097			0.049			0.004
Observations	990	990	990	990	990	990	990	990	990
Subjects	55	55	55	55	55	55	55	55	55
(Pseudo) R^2	0.509	0.534	0.509	0.287	0.328	0.287	0.131	0.140	0.131

Notes. This table shows parameter estimates from OLS regressions (Panel A and B) and average marginal effects from fractional probit response regressions (Panel C). Robust standard errors clustered for subjects are shown in parentheses. P4P is a dummy variable indicating the introduction of P4P. INTERMSEV and HIGHSEV are dummy variables for intermediate and high severities of illness. HIGHMHB is a dummy for the marginal health benefit being 1 if $\theta = 2$ (high), and 0 otherwise ($\theta = 1$, low). All models control for individual characteristics which comprise gender, medical background (non-medical student, medical student, physician), and personality traits; for the respective estimates, see Table B.3.7 in Appendix B.3. * $p < 0.10$, ** $p < 0.05$, and, *** $p < 0.01$.

respectively; see Model (1) of Table 2.3. Nevertheless, the quality of care is significantly lower for intermediately (severely) ill patients by on average 0.44 (0.96) higher deviations from patient-optimal care. Correspondingly, the proportional health benefit is on average 14.3 (14.9) percentage points lower for these patients; see Models (4) and (7) of Table 2.3.

Hypothesis 4 states that under CAP+P4P, the P4P effect increases in the patient's severity of illness and in the marginal health benefit. Our estimations based on Equation (2.4) provide the following results: While the average effect of P4P on the quantity and the quality of medical services is positive and significant for all severity levels, we find substantial heterogeneity when splitting the P4P effect by severities. P4P enhances q by, on average, 0.32, 1.04, and 1.90 medical services for mildly, intermediately, and severely ill patients. This corresponds to a reduction in ρ by, on average, 0.31, 1.31, and 1.91 medical services, and to an increase in \hat{H} by, on average, 5.5, 15.7, and 18.0 percentage points, respectively; Wald tests show that effect sizes are significantly different from each other, see Models (2), (5), and (8) of Table 2.3. Under CAP, the quantity of medical services for severely ill patients deviates the most from the patient-optimal quantity, resulting in the lowest proportional health benefit (Table B.3.1). As physicians respond to P4P, these patients are those who benefit the most from introducing CAP+P4P, which is in line with Hypothesis 4.

We also find that patients with a high marginal health benefit receive significantly more medical services and quality of care compared to patients with a low marginal health benefit. Moreover, while both patient types benefit from CAP+P4P, the patients with a low marginal benefit gain more from the introduction of P4P than those with a high marginal benefit; see Models (6) and (9) of Table 2.3 and the respective Wald tests. This pattern is not in line with Hypothesis 4. However, differences in effect sizes of P4P for patients with a low and high level of marginal health benefit are rather small, and adding interaction terms of marginal health benefits and P4P does not explain better the variation in our data (comparing Models (1) to (3), (4) to (6), and (7) to (9)). In sum, we state:

Result 4 (CAP+P4P and patients' health characteristics) *The effect of performance pay significantly increases in patients' severities of illness. Patients with a low as well as a high level of marginal health benefit gain from performance pay; yet, the effect on quality is*

smaller for patients with a higher marginal benefit.

Results 3 and 4 are in line with findings by Brosig-Koch et al. (2020) for the effect of P4P on treatment quantity and the quality of care. In their study, the effects for the marginal health benefit go in the same direction, but they are statistically not significant.

2.4.4 Comparison of performance pay effects in capitation and fee-for-service payment systems

In this section, we investigate how subjects' responses to performance pay differ between FFS and CAP conditions. Although the payment systems are structurally symmetric due to our mirror-image design, the effect sizes may be different for the following reasons. FFS, a piece-rate system with fees higher than marginal costs, incentivizes the provision of care to be more than patient-optimal. Under CAP, however, physicians have an incentive for underprovision as each medical service provided is costly, thus reducing the physician's profit. Depending on the baseline payment condition, introducing P4P provides incentives that go in opposite directions: to reduce services under FFS and to expand treatment under CAP.

To address Hypothesis 5, we investigate whether the severity-specific effects of P4P on the quality of care differ between FFS+P4P and CAP+P4P.²⁹ Figure 2.2 shows that the effects of blended P4P systems on ρ seem to depend on the patient's severity of illness. In order to investigate Hypothesis 5 further, we use regression analyses, extending our basic econometric model by the between-payment system comparison as follows:

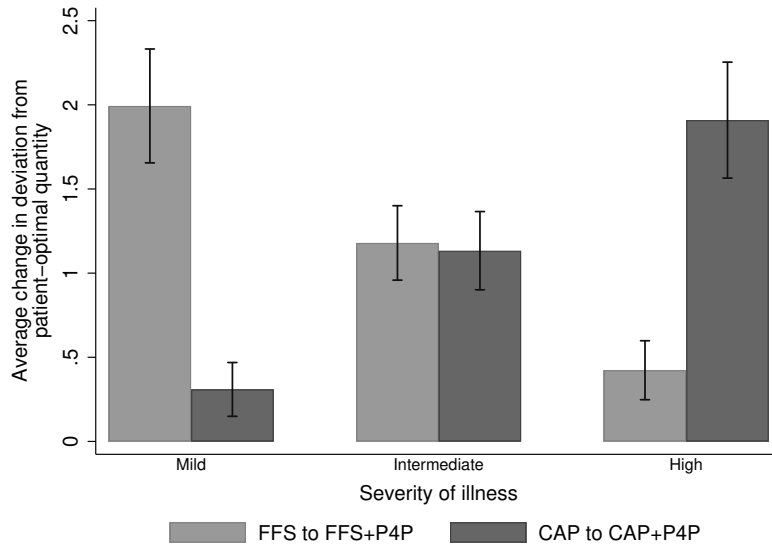
$$\begin{aligned}
y_{ij} = & \alpha + \beta_1 \text{CAP} + \beta_2 \text{INTERMSEV} + \beta_3 \text{HIGHSEV} + \beta_4 \text{HIGHMHB} \\
& + \beta_5 \text{CAP} \times \text{INTERMSEV} + \beta_6 \text{CAP} \times \text{HIGHSEV} \\
& + \beta_7 \text{CAP} + \text{P4P} \times \text{MILDSEV} + \beta_8 \text{FFS} + \text{P4P} \times \text{MILDSEV} \\
& + \beta_9 \text{CAP} + \text{P4P} \times \text{INTERMSEV} + \beta_{10} \text{FFS} + \text{P4P} \times \text{INTERMSEV} \\
& + \beta_{11} \text{CAP} + \text{P4P} \times \text{HIGHSEV} + \beta_{12} \text{FFS} + \text{P4P} \times \text{HIGHSEV} \\
& + \beta_{13} \mathbf{X}_i + \epsilon_{ij}.
\end{aligned} \tag{2.5}$$

²⁹Note that a comparison of the effect differences in quantity appears unreasonable, since P4P leads to opposite responses, i.e., a decrease under FFS or an increase under CAP. Thus, a joint comparison of effects on quantity of care for both payment systems does not allow us to draw any meaningful inferences.

The variable CAP is a dummy which equals 1 if a physician is remunerated by CAP, and 0 if he or she is remunerated by FFS. $\text{INTERMSEV} \times \text{CAP}$ and $\text{HIGHSEV} \times \text{CAP}$ show interaction effects between CAP and the respective level of severity. To determine how severity-specific effects of P4P vary by the underlying remuneration condition, we interact the variables CAP+P4P and FFS+P4P, which are dummies for the respective blended payment systems with each level of severity. The estimate for β_7 thus represents the total effect of P4P for mildly ill patients under CAP, while β_8 represents the total effect for mildly ill patients under FFS and, respectively, for β_9 to β_{12} .

Table 2.4 provides more detailed descriptive statistics on the two quality measures ρ

Figure 2.2: Reduction in the absolute deviation from optimal care by payment system and severity of illness



Notes. This figure shows the reduction in ρ achieved by introducing performance pay, differentiated by FFS and CAP conditions and severities of illness.

and \hat{H} differentiated by patients' severities of illness and marginal health benefits. For mildly ill patients, the improvement in the quality of care is significantly higher under FFS+P4P than under CAP+P4P ($p < 0.001$, two-sided Mann-Whitney U-tests for both quality measures). We observe the reverse pattern for severely ill patients ($p < 0.001$) and no significant differences for patients with an intermediate severity of illness ($p \geq 0.598$). When differentiating patients by their marginal health benefit, no significant differences in

the P4P effect across payment conditions are found; neither for patients with a low nor with a high level ($p \geq 0.308$).³⁰

Table 2.4: Comparison of effects of performance pay blended with fee-for-service and capitation on the quality of care

	FFS to FFS+P4P	CAP to CAP+P4P	Diff.	p -value
A. Change in absolute deviation from optimal care ρ				
Aggregate	-1.20 (1.73)	-1.12 (1.79)	-0.08	0.278
Mild severity	-1.99 (2.14)	-0.31 (1.04)	-1.68	<0.001
Intermediate severity	-1.18 (1.40)	-1.13 (1.51)	-0.05	0.598
High severity	-0.42 (1.11)	-1.91 (2.24)	-1.49	<0.001
Low marginal health benefit	-1.21 (1.78)	-1.17 (1.87)	-0.04	0.519
High marginal health benefit	-1.17 (1.63)	-1.01 (1.63)	-0.16	0.316
B. Change in proportional health benefit \hat{H}				
Aggregate	0.19 (0.27)	0.18 (0.29)	0.01	0.286
Mild severity	0.28 (0.31)	0.04 (0.15)	0.24	<0.001
Intermediate severity	0.24 (0.28)	0.23 (0.31)	0.01	0.608
High severity	0.06 (0.16)	0.27 (0.32)	0.21	<0.001
Low marginal health benefit	0.19 (0.28)	0.19 (0.30)	0.00	0.550
High marginal health benefit	0.19 (0.26)	0.16 (0.26)	0.03	0.308
Observations	468	495		
Subjects	52	55		

Notes. The table reports descriptive statistics on the changes in our quality measures ρ and \hat{H} when moving from unblended to pay for performance payment schemes (means; standard deviations in parentheses). We differentiate by patients' severities of illness and the marginal health benefit. Column "Diff" reports average differences in effect sizes between both payment schemes; reported p -values are based on two-sided Mann-Whitney U tests.

Table 2.5 presents estimates for two versions of Equation 2.5 for each quality measure which differ in that they include X_i as the vector of subject i 's characteristics. Effect differences at a between-subject level may be sensitive to individual characteristics. We find that our estimates on the severity-specific effects of P4P are robust to controlling for individual characteristics (comparing Models (1) to (2) and (3) to (4) of Table 2.5). For simplicity, we, thus focus on Models (2) and (4) when describing our estimation results.

Our findings that the effects of blended P4P systems are severity-specific support

³⁰For the effect of patients' marginal health benefits, see the estimation results in Table B.3.8 in Appendix B.3.

Hypothesis 5. We find that the marginal benefit of P4P on the quality of care is highest for mildly ill patients under FFS+P4P. Models (2) and (4) of Table 2.5 show that the absolute deviation from the patient-optimal quantity is reduced by on average 1.99 medical services, and the patients' health benefit increases by about 17.0 percentage points. On the contrary, the effect is lowest for mildly ill patients under CAP+P4P. Estimates indicate a reduction in ρ by about 0.31 medical services and an increase in \hat{H} by 5.4 percentage points. The introduction of P4P is therefore 6.5 times (3.1 times) more effective in terms of ρ (\hat{H}) for mildly ill patients under FFS+P4P than under CAP+P4P.

For severely ill patients, the estimates show a reverse pattern in that the P4P effect is significantly higher under CAP+P4P compared to FFS+P4P. P4P leads to a reduction in ρ by about average 1.91 medical services under CAP+P4P and about 0.42 medical services under FFS+P4P. \hat{H} increased by 16.7 (7.7) percentage points under CAP+P4P (FFS+P4P). For intermediately ill patients, we find no significant difference in P4P effects between payment systems. Put differently, the introduction of P4P yields similar quality improvements for intermediately ill patients, which lead to a reduction of about 1.13 (1.18) in ρ and a higher \hat{H} by about 14.8 (15.0) percentage points under CAP+P4P (FFS+P4P). In sum, we state the following result:

Result 5 (Comparisons of FFS+P4P and CAP+P4P) *The effect of performance pay on the quality of care is specific to the patient's severity of illness for the two blended pay for performance systems. While the effect on quality of care is significantly higher for mildly ill patients under FFS+P4P, it is significantly higher for severely ill patients under CAP+P4P. For intermediately ill patients the effect of performance pay on the quality of care does not differ between payment systems.*

Table 2.5: Comparison of effects of blended performance pay systems

Method: Model:	A. Absolute deviation from patient-optimal care ρ		B. Proportional health benefit \hat{H}	
	OLS (1)	OLS (2)	Frac. Probit (3)	Frac. Probit (4)
CAP	-1.828*** (0.342)	-1.832*** (0.310)	0.210*** (0.037)	0.209*** (0.032)
INTERMSEV	-0.936*** (0.147)	-0.936*** (0.148)	0.020* (0.012)	0.020 (0.012)
HIGHSEV	-1.782*** (0.253)	-1.782*** (0.254)	0.182*** (0.021)	0.178*** (0.020)
HIGHMHB	-0.086** (0.033)	-0.086** (0.034)	0.013** (0.005)	0.013** (0.005)
CAP \times INTERMSEV	1.784*** (0.201)	1.784*** (0.201)	-0.243*** (0.029)	-0.241*** (0.028)
CAP \times HIGHSEV	3.540*** (0.354)	3.540*** (0.355)	-0.492*** (0.033)	-0.484*** (0.033)
CAP + P4P \times MILDSEV	-0.309*** (0.107)	-0.309*** (0.107)	0.056*** (0.017)	0.054*** (0.016)
FFS + P4P \times MILDSEV	-1.994*** (0.274)	-1.994*** (0.275)	0.171*** (0.017)	0.170*** (0.016)
CAP + P4P \times INTERMSEV	-1.133*** (0.187)	-1.133*** (0.188)	0.148*** (0.018)	0.148*** (0.017)
FFS + P4P \times INTERMSEV	-1.179*** (0.181)	-1.179*** (0.182)	0.151*** (0.017)	0.150*** (0.016)
CAP + P4P \times HIGHSEV	-1.909*** (0.289)	-1.909*** (0.290)	0.168*** (0.018)	0.167*** (0.017)
FFS + P4P \times HIGHSEV	-0.423*** (0.110)	-0.423*** (0.111)	0.075*** (0.015)	0.077*** (0.015)
Constant	2.759*** (0.316)	2.950*** (0.324)		
Individual controls	No	Yes	No	Yes
Wald tests (p -value):				
H_0 : CAP + P4P \times MILDSEV = FFS + P4P \times MILDSEV	<0.001	<0.001	<0.001	<0.001
H_0 : CAP + P4P \times INTERMSEV = FFS + P4P \times INTERMSEV	0.860	0.860	0.872	0.884
H_0 : CAP + P4P \times HIGHSEV = FFS + P4P \times HIGHSEV	<0.001	<0.001	<0.001	<0.001
Observations	1926	1926	1926	1926
Subjects	107	107	107	107
(Pseudo) R^2	0.240	0.312	0.094	0.129

Notes. For Panel A, OLS estimates are reported with robust standard errors clustered for subjects (in brackets). For Panel B, average marginal effects based on a fractional probit response model are reported with robust standard errors clustered for subjects (in brackets). CAP = 1 if physicians are remunerated by CAP, and = 0 otherwise (by FFS). P4P is a dummy variable indicating the introduction of P4P. INTERMSEV and HIGHSEV are dummy variables for intermediate and high severities of illness. HIGHMHB is a dummy for the marginal health benefit being 1 if $\theta = 2$ (high), and 0 otherwise ($\theta = 1$, low). Controls for subjects' individual characteristics comprise gender, medical background (non-medical student, medical student, physician), and personality traits; for the respective estimates, see Table B.3.9 in Appendix B.3. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

2.5 Discussion

2.5.1 Implications

In this section, we discuss implications of our behavioral results from a health policy perspective. In particular, we focus on whether and, if so, under which circumstances a P4P system with discrete bonuses and quality thresholds based on health outcomes is effective to improving the quality of care. Needless to say, we thereby keep in mind that the behavioral results need to be interpreted within the confines of our experimental design and parameterization (for a discussion of limitations, see Section 2.5.3).

Our behavioral results suggest that the effects of P4P on the quality of care are specific to the baseline payment system FFS or CAP, and the patients' severity of illness. The largest effects occur when physicians' financial interests (profit) and patients' interests (health benefit) are most misaligned under FFS and CAP. This means that mildly ill patients benefit the most from P4P when it is blended with FFS, whereas severely ill patients benefit most when it is blended with CAP. For intermediately ill patients, P4P-induced improvements in the quality of care are comparable across payment systems. These results seem to encourage the introduction of P4P to counteract misaligned financial incentives for overprovision and underprovision under FFS and CAP, respectively. When designing P4P systems, health policy-makers should take into account that P4P effects are specific to the patients' health characteristics and the underlying baseline remuneration.

Identifying and classifying patient types based on their severities (and needs of medical services) might be complicated for designing performance-based thresholds in the real world. While it might be possible to discriminate between the different severity of illness types in general, a more precise distinction for designing P4P bonus payments with a sufficient precision might not be feasible or might require excessively high screening costs. In this case, a change in the baseline payment system from CAP to FFS and vice versa could provide an alternative to introducing P4P. We address these considerations by testing for equality of the estimated coefficients of Model (3) of Table 2.5. First, we test whether the absolute deviation from patient-optimal care for mildly ill patients significantly differs if treated under CAP

or FFS+P4P. We find no significant differences in coefficients ($p = 0.272$). Second, we test whether the quality of care for severely ill patients under FFS differs from CAP+P4P. Again, we find no significant differences in the quality of medical service provision ($p = 0.278$).³¹

As a key implication of these findings, a change of the baseline payment systems for mildly and severely ill patients seems to be a meaningful alternative to the introduction of P4P to enhance the quality of care. For intermediately ill patients, however, a significant increase in the quality of care can only be achieved by introducing P4P regardless of the baseline payment condition. To ensure a fairly high quality of medical care, mildly and intermediately ill patients should be treated under a CAP baseline payment, ideally blended with precisely designed P4P bonus payments. Even if we assume that imprecisely designed P4P bonus payments do not lead to any improvements in the quality of care compared to the non-blended condition, the quality of care for these patients would still be better under (non-blended) CAP than under FFS. For severely ill patients, on the other hand, potentially misaligned P4P components could lead to a fairly low quality of medical care under CAP, as a non-blended CAP system turns out to be the most harmful for these patients with highest deviations from patient-optimal care. Thus, a (non-blended) FFS system seems more reasonable to ensure a reasonably high quality of medical care for severely ill patients. Moreover, this would also prevent severely ill patients from possibly not being treated under CAP in the first place due to cream-skimming of healthier patients (e.g., Iversen and Lurås, 2000; Glazier et al., 2009).

This implication is in line with considerations by Barham and Milliken (2015), who show that severely ill patients should be treated by physicians paid by FFS, whereas the physicians treating rather healthy patients should be paid by CAP. Recent empirical evidence by Chami

³¹Based on the estimated coefficients of Model (2) in Table 2.5, we calculated the respective absolute deviations of medical services from the patient-optimal quantity. First, we compare the absolute deviation for mildly ill patients treated under CAP to those treated under FFS+P4P. We find that the quantity of medical services deviates on average about 1.12 medical services from the patient-optimal quantity for mildly ill patients under CAP ($\text{CONSTANT} + \text{CAP} = 2.95 + (-1.83) = 1.12$), while it deviates by on average 0.96 for mildly ill patients under FFS+P4P ($\text{CONSTANT} + \text{FFS} + \text{P4P} \times \text{MILDSEV} = 2.95 + (-1.99) = 0.96$). Based on a Wald test, we cannot reject the null hypothesis that the coefficients are equal ($p = 0.272$). Second, we calculate the coefficients for severely ill patients under FFS and under CAP+P4P. On average, it deviates by 1.17 for highly ill patients under FFS ($\text{CONSTANT} + \text{HIGHSEV} = 2.95 + (-1.78) = 1.17$) compared to 0.97 for severely ill patients under CAP+P4P ($\text{CONSTANT} + \text{CAP} + \text{HIGHSEV} + \text{CAP} \times \text{HIGHSEV} + \text{CAP} + \text{P4P} \times \text{HIGHSEV} = 2.95 + (-1.83) + (-1.78) + 3.54 + (-1.91) = 0.97$). Again, we cannot reject the null hypothesis when testing for equality of coefficients ($p = 0.278$).

and Sweetman (2019) on primary care physicians (who generally treat a rather healthy population) reveals that changing the remuneration from FFS+P4P to a baseline CAP payment system, which is mixed with FFS for non-basic services, leads to a reduction in inappropriate medical service provision. The introduction of mixed FFS and CAP payment systems (CAP+FFS) complements the consideration that changing the underlying payment system might serve as a meaningful alternative to improve quality of care. This argument is also supported by findings of Brosig-Koch et al. (2017a).

2.5.2 Benefits and costs of introducing performance pay

Most research on the effect of P4P initiatives has focused on the targeted quality measures, thereby often neglecting the pertinent issue of their effect on health outcomes and costs (Meacock et al., 2014). In the following, we address this issue by analyzing the effect of introducing P4P on health outcomes (measured as patients' health benefit) and costs of incentive payments.³²

Given our experimental parameters, the average patient health benefit \bar{H} is 7.92 and 8.01 in FFS and CAP, respectively (see Table 2.6). Under P4P, \bar{H} significantly increases to 9.47 in FFS+P4P and to 9.51 in CAP+P4P ($p < 0.001$, Wilcoxon signed rank-test).³³ Also, the physicians' remuneration increases significantly when introducing P4P compared to the non-blended payment schemes ($p < 0.001$, Wilcoxon signed rank-test). This finding is in line with, for example, Mullen et al. (2010) and does not come as a surprise, as subjects in our experiment do react to the more costly P4P incentives.

Moreover, it has been argued that the key to an effective P4P system is the design of its elements (Epstein, 2012; Maynard, 2012; Kristensen et al., 2016). Therefore, we take a closer look at cost and benefits for different severities of illness, as physicians' incentive payments are systematically varied for the three different severities of illness. First of all,

³²Notice that, for a cost effectiveness analysis of P4P schemes, Meacock et al. (2014) propose the following "cost categories": (i) set up/development costs (e.g., staff time, infrastructure investment); (ii) running costs (e.g., administration); (iii) incentive payments, (iv) costs to providers of participating in the scheme, (v) cost savings (e.g., reduced complications, length of stay, readmissions) due to improving the quality of care resulting in superior health outcomes. In our parsimonious experimental design, we focus on the effect of additional *incentive payments* and therefore also restrict our analysis of costs to this category.

³³In absolute terms, the maximum health benefit which can be achieved by patient-optimal medical service provision is on average 10.33.

Table 2.6: Patients' benefits, costs for physicians' remuneration, and changes in costs and benefits

	Aggregated		Mild severity		Interm. severity		High severity	
	\bar{H}	\bar{R}	\bar{H}	\bar{R}	\bar{H}	\bar{R}	\bar{H}	\bar{R}
FFS	7.92	13.38	6.72	11.38	7.92	13.38	9.10	15.38
FFS+P4P	9.51	15.02	9.35	12.93	9.52	14.75	9.65	17.38
Change	1.59	1.64	2.63	1.55	1.60	1.37	0.55	2.00
Ratio ($\Delta R/\Delta H$)	1.03		0.59		0.86		3.64	
CAP	8.01	10.00	9.15	10.00	8.03	10.00	6.86	10.00
CAP+P4P	9.47	13.77	9.53	12.31	9.51	13.53	9.36	15.46
Change	1.46	3.77	0.38	2.31	1.48	3.53	2.50	5.46
Ratio ($\Delta R/\Delta H$)	2.58		6.08		2.39		2.18	

Notes. This table shows the average patients' health benefits \bar{H} and remuneration \bar{R} for FFS, CAP, FFS+P4P, and CAP+P4P, both aggregated and differentiated for severities of illness (mild, intermediate, high). It further shows the marginal payment, marginal patient health benefit, and the ratio of marginal payment to marginal patient health benefit, also aggregated and separately for the three severities of illness.

we find that patients' health benefits and physicians' remunerations (significantly) increase for all severities ($p < 0.010$, Wilcoxon signed rank-test). Under CAP, the increase in health benefit is highest for the severely ill patients (43.7%), while under FFS the increase is highest for mildly ill patients (39.1%). Accordingly, the change in remuneration is 54.6% for the severely ill patients under CAP and 13.6% for the mildly ill patients in FFS. The difference in percentage changes between payment systems indicates that the effectiveness of P4P might vary when costs are also taken into account.

To investigate this further, we now consider the ratio between remuneration and patient health benefits between non-blended and blended P4P payment systems. For a one-unit increase in patient health benefit, physicians' remuneration needs to increase by 2.58 units in CAP conditions and by 1.03 monetary units in FFS conditions. Under CAP, the ratio is lowest for severely ill patients (2.18), due to the large increase in patient health benefit. The ratio is highest for mildly ill patients (6.08), driven by the rather small increase of 4.2% in patient health benefit. Under FFS, the ratio is 0.59 for patients with a mild severity of illness, and for patients with an intermediate severity of illness the ratio is 0.86. This implies an increase in remuneration by less than one monetary unit for an increase in patients' health benefit by one unit for patients with mild and intermediate severity of illness. For severely ill patients, the introduction of P4P is less effective, requiring substantial financial resources (indicated by a ratio of 3.64).

Needless to say, the conversion of one monetary unit to a one-unit change in a patient's health benefit is a stylized simplification. Keeping the limitation of laboratory experiments in mind when interpreting our results beyond qualitative implications, the calculated ratios of marginal payment to marginal patient health benefit can serve as rough guidance for the effectiveness of introducing P4P to the different baseline payment systems. Our results suggest that incentivizing physicians' medical service provision with P4P is a good idea in general for health care policy-makers, whose aim is solely to enhance the patients' health benefit, regardless of the costs due to additional payment to physicians. Taken at face value, it would be most effective to introduce P4P for mildly ill patients under FFS and for severely ill patients under CAP.

Taking into account, however, the relative effectiveness when changing the baseline payment systems instead of introducing P4P, we found the ratio of marginal payment to marginal patient health benefit to be 0.58 when switching from FFS to CAP for mildly ill patients, and 2.40 when switching from CAP to FFS for highly ill patients. Hence, the ratios based on a change in the underlying payment system are close to the ratios based on the introduction of P4P, see Table 2.6. Whereas this suggests that both policy instruments are similar in terms of cost-benefit effectiveness, the highest (absolute) patient's health benefit can only be achieved by a performance-based bonus payment. Keeping in mind that patients' severity types might be difficult to observe in the real world, the introduction of P4P is generally favorable as it leads, at the aggregate, to an increase in patients' benefits for all severity types.

2.5.3 Limitations

Our study faces potential limitations. First, our findings might be sensitive to the design of performance-based bonus payments. As briefly outlined in Section 2.5.1, it might be relatively difficult to determine the individual patient-optimal treatment quantity for health care policy-makers in a real-world environment in the absence of a highly controlled setting. Thus, performance-based bonus components according to the patients' specific medical needs most probably are difficult to design. In our study, bonus payments are tied to the severity-specific patient-optimal treatment quantity and also differ in size with regard to the

level of severity. While the latter design element ensures that we can compare effects across payment schemes, we leave it to further research to investigate how behavioral responses might differ with a stable bonus level and thus lower financial incentives for decreasing non-optimal medical service provision. Even though we incorporate information asymmetry in assuming that the quality is not fully contractible and bonus payments are also granted when deviating one unit from the patient-optimal quantity, we do not consider a full lack of information when designing P4P. As a result, our paper does not yield insights on P4P effects in which the appropriate level of medical service provision is uncertain. There is good reason to believe that uncertainty in the effect on the patients' health benefits does not significantly affect medical service provision (indicated by experiment of Martinsson and Persson, 2019). An extension could thus focus on the P4P effect when not only patients' benefit but also payments are somewhat uncertain.

Second, the obtained results are based on stylized laboratory and artefactual field experiments which comprise a highly controlled, but at the same time artificial, decision environment. While the high control of the decision environment points to a weakness, as it might reduce the external validity of the results, it simultaneously represents one of the key advantages of laboratory experiments. In particular, the environment which can be controlled more tightly than in any other context allows us to identify causal effects. In the field, existing institutions are adopted endogenously, which leads to difficulties in rendering causal inferences about their effects. The lab, however, allows exogenous changes in institutions. It provides the opportunities for controlled *ceteris paribus* variations which are necessary to establish a causal link beyond the identification of simple correlations. According to Falk and Heckman (2009): "Laboratory experiments are very powerful whenever tight control [...] is essential. [...] Tight control [...] also allows replicability of results, which is generally more difficult with field data" (p. 537). As we are primarily interested in establishing the causal link between P4P and provision behavior and in comparing experimental data to theoretical predictions, employing a laboratory experiment seems a natural but complementary approach to other methods of analysis. While taking into account the objections towards lab experiments, we cautiously derive direct external implications from our experimental data; see also the discussion in Hennig-Schmidt et al. (2011).

Conducting laboratory experiments can be considered a complementary rather than a substitute to other methods. Lab experiments can serve as a test bed for investigating the behavioral impact of policy interventions like introducing P4P in our case, before they are implemented in the field. It provides insights into behavioral responses to policy interventions also in health care policy, like varying physicians' incentives, without adverse effects for patients' health. While we do not claim that results from laboratory testing always readily apply to real-world decision environments, the ability to test behavioral hypotheses under carefully controlled conditions provides an indispensable (and relatively inexpensive) tool to to understand behavioral mechanisms better, before running costly RCTs, for instance, or directly implementing policy measures in the field (Galizzi and Wiesen, 2018). Falk and Heckman (2009) argue that economic engineering, a combination of theory and experiments, has improved the design and functioning of markets and institutions. Examples in health care are matching doctors to entry level positions in the general medical labor market (Roth, 2002) or testing clinical decision support systems to improve decisions about hospital discharges (Cox et al., 2016a).

Finally, recent study have shown that the criticism of lab experiments having low external validity may be not as serious as it might appear at first sight. Herbst and Mas (2015) published a direct systematic comparison of the same economic parameter estimated in laboratory experiments and field studies - the estimated spillover effect of worker productivity on the productivity of co-workers - based on 39 studies. The authors show that the results of this class of lab experiments can be generalized to the field because they provide quantitatively precise descriptions of productivity spillovers between workers (Charness and Fehr, 2015). Running a large-scale incentivized survey, a very recent study by (Snowberg and Yariv, 2021) find that different types of elicited behavioral attributes, e.g., risk-aversion, altruism, over-confidence, over-precision, implicit attitudes toward gender and race, various strategic interactions, are comparable between undergraduate university students, a representative sample of the general US population and a sample of Amazon Mechanical Turk, a crowdfunding marketplace commonly used to conduct large-scale economic studies.

2.6 Conclusion

While the idea of using P4P for physicians as a way of improving health care outcomes has increasingly made its way into health policy, the effects on physicians’ provision behavior and patients’ health benefits are not well understood. To this end, we performed controlled laboratory and artefactual field experiments to analyze the causal effect of pay for performance on medical service provision. At a within-subject level, we introduced P4P - with performance thresholds tied to the patient-optimal treatment and adjusted for the patients’ severity of illness levels - which either complements FFS or CAP. Under P4P, subjects increase, on aggregate, the quality of health care provision compared to non-blended payments. We unpack the positive effect of P4P, as we find that the intensity of an effect is significantly driven by the patients’ severity of illness. At a between-subjects level, we are able to determine further how the severity-specific behavioral responses to P4P differ between both baseline payment systems. While the quality of medical services for intermediately ill patients increases likewise by the introduction of P4P under both payment systems, mildly ill patients marginally benefit the most from P4P when complementing FFS, as opposed to highly ill patients, who benefit the most of the introduction of P4P when complementing CAP.

In our parsimonious experimental design, we reduced the complexity of a physician’s treatment decisions, abstracted from multitasking, considered a one-dimensional quality, and we refrained from measurement issues of a physician’s quality of treatment. In contrast, we focused on *exogenously* introducing P4P while keeping all other variables constant. We incentivized physicians for certain health outcomes - in particular, if a physician’s treatment choice either renders the patient’s health benefit or deviates only by one unit from the patient-optimal treatment - which neither generated uncertainty in physicians’ payoffs nor in patients’ outcomes. Taking a more general perspective, a controlled lab experiment could be regarded as a “wind tunnel study”, which allows us to test for the behavioral effects of important design elements of P4P prior to implementing these elements, for example in a large-scale randomized controlled trial (RCT) in the field (Galizzi and Wiesen, 2018).

In our experiment, we found P4P implying an increase in a physician’s maximum

attainable payoffs by 20% to be effective in inducing a higher quality of medical service provision. Also, our behavioral data showed that adjusting P4P for the patients' severity of illness was reasonable to cope with strong overtreatment of low-severity patients under FFS and with strong undertreatment of high-severity patients under CAP. P4P bonus payments are designed such that performance thresholds which are tied to the patient-optimal care are accurately adjusted for severity of illness levels and account for severity-specific patients' benefits by precisely varied bonus sizes. It might not always be feasible to design P4P bonus payments such accurately. Whenever this is the case, a general distinction between patient groups of rather high and low medical needs, however, is possible, patients belonging to the former group should be always be treated under FFS, while patients with knowingly rather low medical needs should be treated under CAP. This approach would guarantee that harm to patients, which is caused by deviations from the patient-optimal medical care induced by opposing financial incentives between physician profit and patient benefit, is kept small.

Further, cost effectiveness is another important issue for the economic evaluation of P4P systems (Meacock et al., 2014). It is often argued that P4P schemes can be considered cost-effective when quality increases are achieved with equal or lower costs or when the same quality is achieved with lower costs. When quality improvements are large, P4P may be viewed as cost-effective even if it leads to cost increases. In general, however, the evidence on cost effectiveness of P4P seems rare and inconclusive. Most of the studies suggest quality improvements at the expense of cost (Emmert et al., 2012; Meacock et al., 2014). Analyses of effectiveness based on our experimental data reveal that the patients' health benefit increases through P4P. On the other hand, additional expenditures for physicians' payments rise disproportionately. Note that our calculations are limited to incentive costs. However, there are other "cost categories" which might be affected by the introduction of P4P. For example, "cost savings" seem likely as P4P induces care with superior health outcomes, which in turn has consequences for future health care costs. A health care policy-maker might also take these considerations into account when evaluating the (cost-)effectiveness of P4P schemes.

Finally, physicians respond differently to incentives from P4P. This calls for future work to understand better what drives the heterogeneity. What is the role, for example, of individuals'

underlying social preferences or personality traits? These individual characteristics might not only explain health care workers' responses to performance pay (e.g., Donato et al., 2017) and self-selection into the payment systems (e.g., Dohmen and Falk, 2011; Brosig-Koch et al., 2017b). Knowledge on the underlying preferences that predict sorting are therefore of great importance for researchers and health care policy-makers alike. Moreover, policy-makers need to account for physicians' capabilities to manipulate the systems by only treating patients for whom they are able to generate the performance-based bonus payments as an unintended consequence of incentive schemes emphasized by a recent study evaluating a pilot program on reducing hospital costs (Alexander, 2020). Finally, further research is needed to compare systematically the effects of different blended payment menus of payment systems, as opposed to the introduction of quality-based P4P with respect to costs, benefits, and feasibility.

Chapter 3

Physicians' performance pay and personality traits

3.1 Introduction

A key question in health economics addresses the issue on how to remunerate physicians to improve the quality of care. It is widely acknowledged that traditional approaches to remunerate physicians provide financial incentives to deviate from a patient-optimal level of care. Fee-for-service (FFS) payments encourages overprovision and lump-sum capitation (CAP) payments underprovision of medical services. Pay for performance (P4P) bonus payments present a promising policy approach to align physicians' financial incentives with a high quality of care.

Despite the increasing popularity of P4P initiatives in practice³⁴, empirical evidence on the effect on the quality of care is rather mixed and inclusive (for extensive literature reviews, see Scott et al., 2011; Mathes et al., 2019; and Jia et al., 2021). Empirical studies typically assess the effectiveness of P4P programs at the aggregate level and face the challenge to identify a causal effect of performance pay on the quality of care. Due to various confounds in the field, health outcomes are difficult to observe and may be biased (Campbell et al., 2009; Gravelle et al., 2010; Roland and Olesen, 2015). Experimental studies represent a valuable complement rendering control over the decision situation, incentive design, and quality measures. While recent experimental evidence shows that introducing performance pay increases the quality of medical care, it also reveals a lot of heterogeneity in behavior (e.g., Brosig-Koch et al., 2020; Oxholm et al., 2021). Up to now, a profound understanding

³⁴In the USA for example, see, e.g., Rosenthal et al., 2006; Stokes et al., 2018; Song et al., 2019, and in the UK, see, e.g., Roland, 2004; Doran et al., 2006; Roland and Campbell, 2014.

of the underlying mechanisms for the observed heterogeneity is mostly lacking.

The physicians' personal characteristics are a likely source contributing to the heterogeneous responses to P4P. Personality traits are defined as an individual's "relatively stable, consistent, and enduring internal characteristics" (American Psychological Association, nd). There is a rapidly growing general economics literature which acknowledges their potential to explain variations in labor market performance and responses to financial incentives (e.g., Bowles et al., 2001; Almlund et al., 2011; Fulmer and Walker, 2015; Cubel et al., 2016). In health economics contexts, some evidence exists that personality traits relate to the performance of medical service providers (Callen et al., 2018; Eilermann et al., 2019). However, very little is known on how personality traits relate to the effect of P4P on providers' performance. To the best of our knowledge, only one experimental study addresses this issue directly. Using a field experiment observing maternity care providers in rural India, Donato et al. (2017) find differences in providers' responses to performance-based financial incentives according to two personality traits. A comprehensive understanding of how and under which conditions personality traits moderate the effect of P4P is still lacking.

In this study, we aim to narrow this gap in the literature. We investigate whether physicians' personality traits affect how they respond to the introduction of P4P under two traditional baseline payment systems (CAP or FFS). These payment schemes differ by inherent financial incentives for either underprovision or overprovision of medical services. To this end, we utilize data from a series of behavioral experiments in an abstract medical frame and surveys on personality traits with physicians, medical students, and non-medical students. The experimental environment allows to identify the causal effect of P4P on a clean measure of quality of care at the individual subject level. Our study provides valuable insights on how personality traits relate to physicians' responses to P4P, absent confounding effects on patients' health outcome. For two baseline payment systems, we address the following research question: Do certain personality traits indicate why some physicians respond more or less to performance-based financial incentives than others?

3.2 Background

A well-established construct to measure personality is the Big Five personality model.³⁵ Differences in personality at the subordinated level are typically reflected by five primary traits; defined as follows: “extraversion” reflects the orientation of one’s interests and energies toward the outer world of people and things rather than the inner world of subjective experience; “neuroticism” reflects a chronic level of emotional instability and proneness to psychological distress; “openness” reflects the tendency to be open to new aesthetic, cultural, or intellectual experiences; “conscientiousness” reflects the tendency to be organized, responsible, and hardworking; and “agreeableness” reflects the tendency to act in a cooperative, unselfish manner (Dictionary of Psychology of the American Psychological Association).

Although the influences of traits on individuals’ behavior may vary depending on the decision context, personality traits are generally suggested to affect many economic and social outcomes; for an excellent overview on the role of personality traits in economics, see, e.g., Borghans et al. (2008), Almlund et al. (2011). We briefly outline the main findings of the streams in literature most relevant for our study.

In labor market economics, the evidence is rather mixed. While experimental studies share the consistent finding that neuroticism is negatively related to performance in real-effort tasks under piece-rate payment schemes (Müller and Schwieren, 2012; Cubel et al., 2016), the relation to other traits is less uniform. Müller and Schwieren (2012) find that openness negatively relates to individual’s performance, whereas there is positive link between conscientiousness and performance is reported by Cubel et al., 2016. In addition, further (mostly survey) studies suggest that agreeableness is associated with lower labor market outcomes (e.g., Nyhus and Pons, 2005), and extraversion with better labor market outcomes (e.g., Fletcher, 2013).

In health economics, the limited empirical evidence on the impact of personality traits on a provider’s performance is rather mixed with respect to the behavioral relevance of each trait. Callen et al. (2018), for example, show that the performance of health care providers

³⁵For an extensive summary on the history and measurement of Big Five personality dimensions, see John et al. (2008).

in Pakistan is related to their personality traits: physicians higher in conscientiousness are more likely to be present at work at (unannounced) health inspections and physicians higher in all traits but openness are less likely to falsely report their attendance compared to those lower in the respective traits. However, within the frame of a laboratory experiment with medical students and non-medical students Hennig-Schmidt et al. (2019) find no systematic effects between personality traits and dishonest behavior when investigating false reports of birth weights in a neonatal care context.³⁶ Eilermann et al. (2019), in turn, report that conscientiousness affects treatment decisions of pediatricians in an artefactual field experiment: more conscientious pediatricians deviate significantly less from the appropriate therapy length of antibiotic treatment for a series of routine pediatric cases. Since treatment decisions in the experiment are not incentivized, interactions between personality traits and responses to financial incentives cannot be studied. Up to now, Donato et al. (2017) present the only experimental evidence on a link between personality traits and P4P in health care. In a maternity care setting in rural India, they study how providers respond to P4P incentives corresponding to their level of conscientiousness and neuroticism. The provider's performance is assessed by the incidence of adverse health outcome, i.e., the post-partum hemorrhage rates among her patients. The positive effect of P4P on a provider's performance is reported to be less strong for more conscientious and more neurotic providers who, at the same time, are the ones to perform better absent P4P.

We add to this first evidence in a health care context. In a more general approach, we do not refer to one particular health care setting, but utilize data from laboratory experiments in an abstract medical frame. Unlike in observational studies, the closely monitored conditions in laboratory settings allow to precisely measure individual's performance whilst accounting for the baseline payment system and a patient's health characteristics. In addition, we account for the mixed evidence on the behavioral relevance of each trait in the labor and health economics literature, making it a natural approach to explore how each of the five traits relates to individual behavioral, rather than restricting our analyses to certain traits a priori.

³⁶Also Groß et al. (2021) in a complementing paper observe no link between personality traits and dishonest reporting behavior.

3.3 Experiment, Data, and Methods

We utilize experimental data from a medically framed laboratory experiment which was designed to investigate systematically how introducing P4P affects medical service provision under either a baseline CAP or a FFS payment system.³⁷ Overall 107 subjects (20 physicians, 44 medical students, and 43 non-medical students) participated in the experiments between 2011 and 2013.³⁸ In a between-subjects design, subjects were randomly assigned to only one payment condition (CAP or FFS with 55 and 52 subjects, respectively), but experience the variation in performance pay at a within-subject level. Thus, each subject participates in two consecutive experimental parts which only differ in the existence of P4P.

In the role of a physician, subjects are asked to decide on the provision of medical services for nine abstract patients in each part. While patients vary in terms of health characteristics within each part, the characteristics of the patient population are kept constant across both parts. In the first part, subjects decide on medical service provision under traditional CAP or FFS (absent P4P), whereas, in the second part, they decide on it when the baseline payment systems are enhanced with performance pay (CAP+P4P or FFS+P4P). All decisions in the experiment are incentive-compatible, such that they bear real consequences. When deciding on the quantity of provided medical services, a subject determines her own profit and the patient's health benefit simultaneously.³⁹ Even though subjects make their decision on a computer screen and no patients are physically present, real patients outside the lab are affected by the subjects' medical service decisions. In particular, the patient's health benefit realized by the participants' decisions is transferred to a charity which exclusively uses the money for treatment of cataract patients.⁴⁰

Every provided medical service leads to costs for a physician. Under CAP, subjects

³⁷For a detailed description on the experimental design and protocol, see Section 2.3 of Chapter 2.

³⁸The medically framed laboratory experiment with 44 medical students and 43 non-medical students was conducted at the University of Duisburg-Essen, Germany, and the artefactual field experiment with 20 physicians at the Academy for Training and Education of Physicians (Akademie für Ärztliche Fort- und Weiterbildung) in Bad Nauheim, Germany. For a summary of descriptive statistics of our sample, see Table B.1.1 in Appendix B.1.2.

³⁹At the end of the experiment, when all subjects had made their decisions, one decision in each part of the experiment was randomly determined to be relevant for a subject's actual payoff and the patient benefit. For further details on the payment procedure, see Section 2.3 of Chapter 2.

⁴⁰In total, EUR 2,509.40 were transferred to the charity, enabling the treatment of 83 cataract patients.

are paid a fixed lump-sum payment which is, then, reduced by the costs incurring for the provision of the chosen quantity of medical services. Under FFS, subjects receive a (piece-rate) fee per medical service rendered to a patient which is always higher than the incurring costs per service. Hence, FFS provides an incentive for subjects to overserve patients, while CAP provides an incentive to underserve patients. To mitigate the inherent financial incentive in FFS (CAP) to provide too many (few) services, P4P payments introduced in part 2 of the experiment are designed to make the interests of physicians and patients become (more) aligned. To this end, subjects receive bonus payments whenever their medical service provision does not deviate from the patient-optimal quantity by more than one medical service. By definition, the patient-optimal quantity always leads to the highest patient health benefit. Note that all parameters of the experiment are common knowledge and subjects are aware of the patient-optimal quantity which varies for individual patient types. Assuming that the quality is not fully contractible due to information asymmetry, we allow variations from the patient-optimal quantity by one unit. By linking performance pay to the patient's health benefit, we reduce the trade-offs between the maximization of profit and the optimization of the patient's health benefit. However, inherent financial incentives for non-optimal service provision under CAP and FFS are still prevalent. The overall design of payment systems is symmetric in level and marginal incentives, and P4P are set such that incentives are comparable between CAP and FFS.

To elicit personality traits, the short version of the BIG Five Inventory (BFI-10) by Rammstedt and John (2007) is included in a post-experimental questionnaire. The BFI-10 asks subjects to rate their agreement to ten statements on a five point Likert scale, two out of the ten statements, thereby, form one trait.⁴¹

To investigate how P4P-effects on the quality of care relate to personality traits, we apply the following estimation model for each baseline payment system:

$$y_{ij} = \beta_0 + \beta_1 \text{P4P}_j + \beta_2 \text{TRAIT}_i + \beta_3 \text{P4P}_j \mathbf{x} \text{TRAIT}_i + \beta_4 \mathbf{X}_i + \beta_5 \mathbf{Z}_j + \epsilon_{ij}, \quad (3.1)$$

⁴¹For a detailed description of the BFI-10, see Table C.1.1 in Appendix C.1.1.

where y_{ij} is the quality outcome of interest, namely the absolute deviation from patient-optimal care ($|q_{ij} - q_j^*|$). It is measured by the absolute number of medical services that a subject's i chosen quantity (q_{ij}) for a patient j deviates from the patient optimal quantity (q_j^*). A physician provides the highest quality of care whenever she is delivering the patient-optimal quantity of medical services ($y_{ij} = 0$), which, by definition, implies the highest health benefit for a patient. The larger the difference to the patient-optimal quantity, the lower is the quality. $P4P_j$ is a dummy variable indicating the introduction of performance pay. $TRAIT_i$ is a vector of subjects' personality traits representing the score of each personality of the BFI-10. Our main interest are the interaction effects between P4P and each $TRAIT_i$.⁴² X_i represents a vector of time-invariant subject characteristics, i.e., gender and medical experiences (non-medical student, medical student, or physician). Though not our research focus, we account for the impact of patients' health characteristics on quality of care (reported in Chapter 2) by including patient's severity of illness and marginal health benefit per provided service as controls in our subsequent regression analyses, represented by vector Z_j .

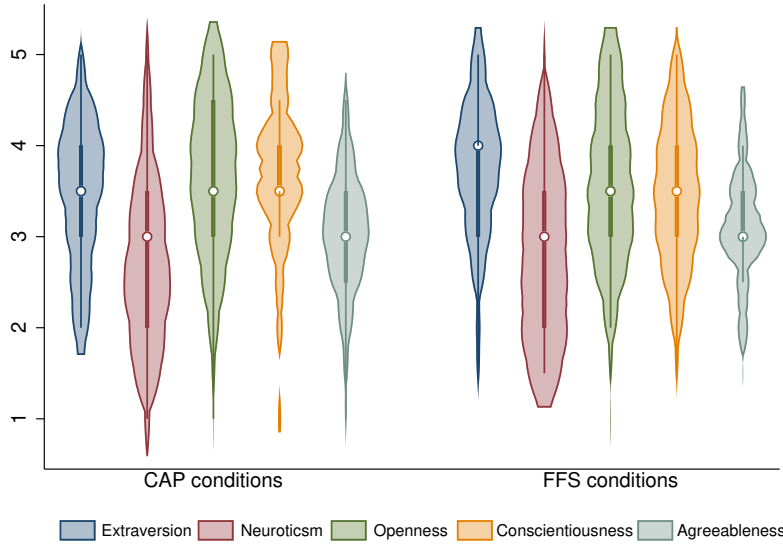
3.4 Results

We now analyze how the effects of P4P on the quality of care interacts with individuals' personality traits.⁴³ Figure 3.1 shows the distributions of personality traits for subjects under each baseline payment system. On average, our subjects exhibit fairly high levels of conscientiousness, extraversion and openness, moderate levels of agreeableness and rather

⁴²To detect a potential problem of multicollinearity when including all traits in a joint model, we checked for several warning signals. Neither the tolerance and variance inflation matrix, nor the correlations of estimated coefficients give reason for concern. We performed separate trait by trait regressions for robustness analyses and report estimates in Tables C.2.3 and C.2.4 in Appendix C.2 for CAP and FFS, respectively. While the findings are qualitative similar to the estimates of our joint model, we note that the adjusted R^2 is higher for our model.

⁴³For a systematic comparison on the overall and patient-severity specific effects of P4P on the quality of care between CAP and FFS, see Chapter 2.

Figure 3.1: Distribution of personality traits among subjects by payment conditions



Notes. This figure shows violin plots on the distributions of personality traits among subjects by payments conditions. Personality traits are measured by the BIG Five Inventory (Rammstedt and John, 2007). Subjects rate their agreement to ten statements on a five point Likert scale with 1 corresponding to disagree strongly, 2 to disagree a little, 3 to neutral, 4 to agree a little, and 5 to agree strongly, two out of the ten statements form one trait.

low levels of neuroticism.⁴⁴ As Figure 3.1 shows, there is a substantial degree of between-subject heterogeneity for each trait which makes it reasonable to study whether variations in personality traits relate to variations in responses to P4P. The variations across CAP and FFS seem similar whatsoever.⁴⁵

We, first, focus on the effect of P4P and personality traits on the quality of care under CAP, applying Equation (3.1). Panel A of Figure 3.2 illustrates our main findings with the regression results reported in Table 3.1.⁴⁶ Recall that quality of care is assessed by

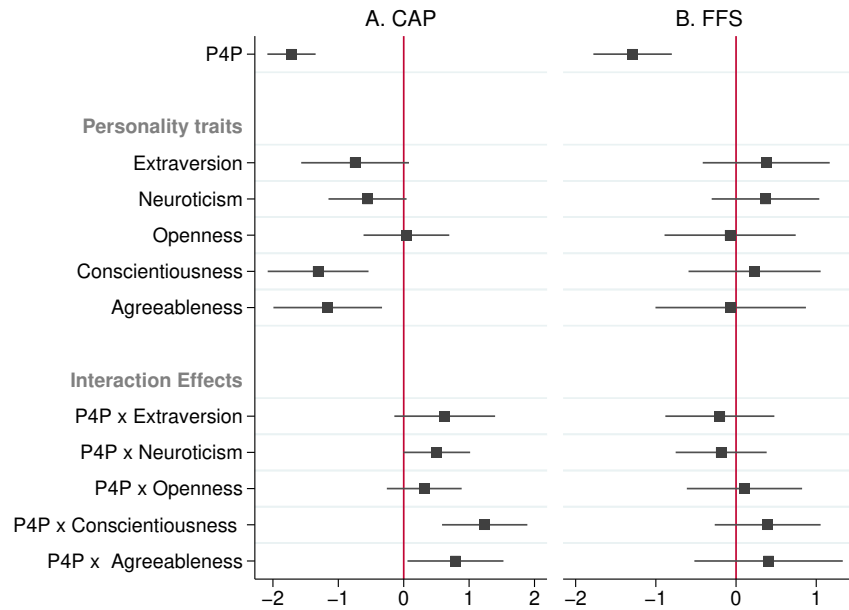
⁴⁴Note that the measurement reliability (Cronbach alpha; α) for our data ranges only from good to fairly low; extraversion ($\alpha = 0.74$), neuroticism ($\alpha = 0.58$) openness ($\alpha = 0.42$), conscientiousness ($\alpha = 0.42$), and agreeableness ($\alpha = 0.18$). Since α coefficients depend on the number of items used, it is not surprising that the consistency is rather low (Furnham, 2008). Thus, when using Cronbach's alpha as reliability measure for a two-item scale, the true reliability is usually underestimated (Eisinga et al., 2013). Our α 's do not appear particularly striking compared to the reported α 's of other studies using the BFI-10, which range from $\alpha = 0.03$ to $\alpha = 0.75$ (e.g., Thalmayer et al., 2011; Crede et al., 2012; Carciofo et al., 2016; Balgiu, 2018 and Hennig-Schmidt et al., 2019).

⁴⁵Overall, samples under CAP conditions and FFS conditions appear fairly similar in all personality measures. P -values of Kolmogorov-Smirnov (KS) tests ($p \geq 0.382$) indicate that there is no statistical evidence of a difference in personality traits between the distributions; for a detailed analysis of differences in personality traits between both subsamples, see Appendix C.1.2.

⁴⁶Though not shown and discussed in detail here, we performed several analyses for robustness checks, reported in Appendix C.2. These include regression estimates for several versions of our base model Equation (3.1) (see Table C.2.5), estimations based on tobit and multilevel regressions (see Table C.2.6), and separate trait by trait analyses (see Tables C.2.3 and C.2.4). The estimation results of each of these additional analyses are qualitatively similar to the results shown in Table 3.1.

the absolute deviation from patient-optimal care with negative regression estimates indicating quality improvements. First, P4P leads to an overall higher quality of care. Second, we find that more conscientious and more agreeable subjects provide a higher quality of care. Third, the positive estimates on the interaction between P4P and conscientiousness is statistically significant different from zero. Similarly, the positive interaction effect of P4P and agreeableness is also of statistical significance. These findings imply that the positive effect of P4P on the quality of care decreases for subjects scoring higher in both traits.

Figure 3.2: Regression estimates: Effect of performance pay and personality traits on quality of care



Notes. This figure shows regression coefficients from separate regressions on the impact of performance pay (P4P) and personality traits on the quality of care for both baseline payment systems, capitation (CAP) in Panel A and fee-for-service (FFS) in Panel B. Error bars represent 90 percent confidence intervals, with standard errors clustered at the individual subject level. For a better interpretation of our regression estimates, we rescaled Big Five scores to $[-1, 1]$. Thus an one unit increase in the respective trait reflects a change from the theoretical minimum to the (neutral) midpoint or from the midpoint to the theoretical maximum). All regressions include subject's individual controls, i.e., gender and medical experience, and patients' individual controls, i.e., severity of illness and marginal health benefit per provided service. Regression estimates corresponding to this figure are reported in Table 3.1.

Or put differently, subjects who are less conscientious and agreeable respond stronger to performance incentives. In contrast to more conscientious and agreeable subjects who provide a closer to patient-optimal quality of care absent P4P incentives, these subjects also offer a higher potential for quality improvements in the first place. Thus, P4P may

Table 3.1: Regression model on the interaction effects of performance pay and personality traits on quality of care

	Absolute deviation from patient-optimal care	
	A. CAP	B. FFS
P4P	-1.716*** (0.220)	-1.290*** (0.292)
EXTRAVERSION	-0.743 (0.492)	0.376 (0.473)
P4P×EXTRAVERSION	0.627 (0.460)	-0.202 (0.406)
NEUROTICISM	-0.553 (0.357)	0.367 (0.400)
P4P×NEUROTICISM	0.506 (0.305)	-0.186 (0.339)
OPENNESS	0.041 (0.392)	-0.074 (0.488)
P4P×OPENNESS	0.314 (0.342)	0.105 (0.429)
CONSCIENTIOUSNESS	-1.309*** (0.461)	0.231 (0.492)
P4P×CONSCIENTIOUSNESS	1.238*** (0.390)	0.394 (0.394)
AGREEABLENESS	-1.162** (0.496)	-0.067 (0.560)
P4P×AGREEABLENESS	0.791* (0.439)	0.406 (0.552)
CONSTANT	1.739*** (0.267)	2.667*** (0.376)
Observations	990	936
Subjects	55	52
R^2	0.328	0.336

Notes. This table shows estimates from OLS regressions on the effects of performance pay (P4P) and personality traits on quality of care under capitation (CAP) or fee-for-service (FFS) baseline payment systems. Robust standard errors clustered for subjects are shown in parentheses. All personality traits are measured on a scale from -1 to +1. The coefficients thus reflect the effect of a one-unit change; for example, a change from the theoretical minimum to the (neutral) midpoint or from the midpoint to the theoretical maximum. All models control for patient's health characteristics which comprise severity of illness and patient's marginal health benefit. Both models include subjects' controls comprising gender and medical experience (non-medical students, medical students, physicians). For the respective estimates, see Table C.2.5 in Appendix C.2. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

not be as effective to improve the quality of care for subjects high in agreeableness and conscientiousness under CAP, for whom the quality of care is already higher in the absence of P4P.

We now study behavioral responses to P4P under FFS and the impact of personality traits, therein, see Panel B of Figure 3.2 and Table 3.1. Consistent with previous findings, introducing P4P leads to an increase in the quality of care. Personality traits, however, do not relate to the provided quality of care in any statistical significant way, neither in the absence nor in the presence of P4P. In contrast to our findings under CAP, individual responses under FFS are not affected by personality traits. This difference suggests that subjects seem to perceive both payment systems differently, despite the mirror image design of payment systems in level, marginal profit and patients' benefit.

Taken together, we find no empirical support that individuals respond stronger to performance incentives according to any personality trait. In contrast, our results (if at all) indicate a reversed relationship such that the effect of P4P is weaker among more conscientious and agreeable individuals under CAP. It needs to be emphasized that the effect still remains positive. This adds to the findings of Donato et al. (2017) who report a similar relation for providers higher in conscientiousness and neuroticism.

3.5 Discussion and conclusion

We study how personality traits relate to the effect of P4P when complementing either CAP or FFS. Thereby, we contribute to the growing literature on how P4P initiatives need to be tailored to efficiently improve the quality of care. To this end, we use data from controlled experiments which allow to obtain a clean measure of quality of care and relate it to individuals' personality traits.

Our behavioral data shows that the underlying baseline payment system is crucial for answering the question whether personality traits influence the effectiveness of P4P. We present experimental evidence that certain personality traits, namely conscientiousness and agreeableness, are behaviorally relevant for the quality of care under CAP conditions but not under FFS conditions. This demonstrates that the perception of medical service

provision under both payment systems might differ. Under CAP, more conscientious and more agreeable subjects, who absent performance incentives provide relatively higher quality of care, respond less strong to the incentives compared to subjects lower in these traits.

In general, CAP provides financial incentives for underprovision as every additionally provided medical service is costly and only payoff decreasing. To obtain higher profits (which occur at the expense of lower patients' benefits), subjects need to follow a "no/low effort"-policy. Individuals higher in conscientiousness, who tend to be more responsible and hardworking, might be more reluctant to adopt such a policy. In contrast, the motivation to provide medical services under FFS with general financial incentives for overprovision is made salient in order to enhance profits. At the same time, the provision of medical services still requires efforts since every provided service is costly for the physician. Under CAP, however, deviations from patient-optimal care do require no such effort and might rather be construed as selfish behavior. More agreeable subjects who tend to act in an unselfish manner might, thus, feel more obligated to provide some kind of sacrifice which, under CAP, can only be achieved in providing (more) medical services at the expense of own profit. The reported link to agreeableness adds to experimental findings derived from neutrally framed dictator games which emphasize that subjects' agreeableness is positively associated with higher giving (e.g., Ben-Ner et al., 2004). Note that, in a broader sense, our CAP-decision task relates to giving in a general dictator game decision task.

The introduction of P4P renders different incentives across baseline payment systems, i.e., to enhance the quantity of medical services under CAP and to reduce it under FFS. While subjects irrespective of personality traits respond similar positively strong to P4P under FFS, the positive effect of P4P on the quality of care is weaker for more conscientious and agreeable subjects under CAP. These subjects already provide a fairly close to patient-optimal quality absent P4P though. On the one hand, they might simply respond less strong to P4P because they have less potential for quality improvements compared to others who behave less patient-regarding absent P4P. On the other hand, in line with the finding that individuals who are rather intrinsically motivated are hardly sensitive to financial incentives (e.g., Lagarde and Blaauw, 2017), these subjects might generally respond less strong to financial incentives such as P4P incentives.

Within the confines of the experiment, the key take-aways of this study are as follows. Certain personality traits of physicians mitigate the effectiveness of P4P. More precisely, introducing P4P to CAP might not effectively improve the quality of care in settings particularly attractive for physicians high in conscientiousness and agreeableness. Even though personality traits (if at all) negatively relate to responses to P4P, the moderating effects on the quality of care are not detrimental for patients. Nevertheless, from a viewpoint of cost efficiency, our findings suggests that whether P4P serves as a means to efficiently improve the quality of care under CAP also depends on the shares of physicians high in conscientiousness and agreeableness. While certain personality traits indicate why some physicians respond less to P4P than others, important to highlight, they also indicate a higher quality of care without additional performance incentives. Thus, selecting more conscientious and agreeable physicians might lead to an improved quality of care and presents an alternative approach to the introduction of P4P in care settings which are characterized by CAP payments. In Germany, for example, primary care could represent such a care setting with CAP as the predominant payment form (e.g., Brosig-Koch et al., 2020).

With empirical evidence reporting that certain personality traits can be linked to occupational choices of future physicians (e.g., Mullola et al., 2018; Bexelius et al., 2016), our study adds to the potential of personality traits within the scope of worker selection strategies. In line with recent claims on selecting workers according to specific characteristics to improve the performance (e.g., Ashraf and Lee., 2016), our findings can be useful in developing targeted interventions aimed at selecting an efficient and patient-oriented medical workforce.

Chapter 4

Physician altruism: The role of medical education

4.1 Introduction

Physician altruism is a key characteristic of behavior in healthcare markets and the notion of the benevolent, sacrificial physician is deeply rooted in medical ethics dating back to the ancient Hippocratic Oath (e.g., Pellegrino, 1987; Beauchamp and Childress, 2001). The modern view in economics is coined by Arrow (1963), who emphasized that a physician's behavior is "supposed to be governed by a concern for the customer's welfare which would not be expected of a salesman" (Arrow, 1963, p. 949). Following Arrow, a large literature showed that physician altruism has important implications, for example, on physicians' responses to incentives (e.g., Ellis and McGuire, 1986; Alexander, 2020), concerns for transparency (e.g., Kolstad, 2013), referrals (e.g., Allard et al., 2011; Liu and Ma, 2013), prescription patterns (e.g., Hellerstein, 1998; Crea et al., 2019), and occupational choices (e.g., Nicholson and Propper, 2011; Li, 2018).

Surprisingly, while altruistic preferences play a key role in modeling physicians' behavior, little is known about their origin, their distribution in the population of future doctors, and their formation during the course of education. In this respect, it is unclear how medical education shapes future physicians' patient-regarding altruistic preferences towards the patients' health in the sense of Arrow (1963). It is thus key to, first, better understand the distribution of altruism, in the population of future doctors, and, second, to identify the effect of medical education on physicians' altruism while accounting for the prevailing heterogeneity in a parsimonious way. In the paper at hand, it is our objective to make an important step in this direction. For this purpose we designed an incentivized behavioral

experiment to study medical students' patient-regarding altruistic preferences.

Our paper builds on several streams of the literature on estimating altruistic preferences among medical students and physicians. A first strand of empirical literature estimates altruism among primary care physicians using their prescription choices. This stream of the literature, originating from Hellerstein (1998), relies on a theoretical framework assuming that both the (indirect) utility of the patient and the insurance expenditures enter the utility function of the physician.⁴⁷ Within this framework, empirical studies compare physicians' marginal utility from patient welfare with their marginal disutility from insurance expenditures (e.g., Hellerstein, 1998; Lundin, 2000; Crea et al., 2019). Making use of prescriptions data on seven different drugs from two Swedish pharmacies in 1992 and 1993, Lundin (2000) estimates a random effects probit model for whether physicians prescribed the branded or generic version of the drugs and finds some support for physicians' altruism: higher coverage decreased (increased) the probability of prescribing a generic (branded) version of a drug. Using a national panel register containing all statins prescriptions in Finland from 2003 to 2010, Crea et al. (2019) estimate the likelihood that physicians prescribe generic versus branded versions of statins as a function of the shares of the difference between what patients have to pay out of their pocket and what is covered by the insurance. Estimated coefficients associated with altruism are nearly zero while, instead, Crea et al. (2019) find strong evidence of habits persistence in prescribing branded drugs.⁴⁸

A second strand of literature focuses on health benefits of patients based on experimental economics methods. Compared to studies using medical prescriptions data, behavioral experiments allow to investigate the nature of patient-regarding altruism at an individual-subject level. This approach is theoretically guided by early formalizations of physicians' behavior by Arrow (1963) and Ellis and McGuire (1986), in which a physician's utility

⁴⁷Hellerstein (1998) assumes that the branded version of the drug is more expensive than the generic version. The model shows that, if the physician places a higher weight on the patient's utility than on insurance expenditures, an increase in the insurance coverage decreases (increases) the likelihood of the generic (branded) prescription. An increase in the insurance coverage, in fact, increases insurance expenditures and decreases patient's expenditures, *ceteris paribus*. As both these variables have a similar effect on the physician's utility, higher insurance coverage leads to a lower probability of generic prescribing when the physician values the utility of the patient more than the insurance expenditure.

⁴⁸In this paper, we focus on physician altruism, and we do not consider studies on other healthcare providers like, for instance, Douven et al. (2019) who analyze the altruistic preferences of mental health workers using a large data set from the Netherlands. For a summary of other examples, see Galizzi et al. (2015).

increases in the patient’s health benefit. A prototypical early example is Hennig-Schmidt et al.’s (2011) medically framed laboratory experiment with a small sample of German medical students. Using data from this experiment, Godager and Wiesen (2013) estimate the marginal rate of substitution between patient benefit and profit as a measure of physician altruism. Their estimation results show patient-regarding altruism with substantial heterogeneity therein. Following Godager and Wiesen (2013), Wang et al. (2020) also estimate the distribution of altruism among 178 Chinese medical students and 99 Chinese physicians and compared it to those 42 German medical students participating in Hennig-Schmidt et al.’s (2011) experiment. Their estimates show that physician altruism is quite similar between Chinese doctors, Chinese medical students, and German medical students.

In a third strand of the literature, altruistic preferences of medical students are elicited experimentally in scenarios with no specific connection to a physician-patient relation. The standard experimental setting is a neutrally framed modified dictator game where altruism is identified by the tradeoff between self-interest and other’s benefit. Following the seminal paper of Andreoni and Miller (2002), preferences over monetary sums are decomposed into two qualitatively different tradeoffs: a first tradeoff between self-interest and other’s benefit, and a second tradeoff between equality and efficiency. In two related papers, Li et al. (2017, 2018) use an online experiment to elicit altruistic preferences of 503 US medical students over distributing monetary sums between themselves and an anonymous other person from the American Life Panel.⁴⁹ Both studies report widely heterogeneous social preferences in terms of their altruism and equality-efficiency tradeoffs. Also, Li et al. (2017) report that medical students are similar in altruism, equality and efficiency preferences compared to non-medical student subjects in comparable samples but are substantially less altruistic and more efficiency-focused than the full ALP sample.

In this paper, we designed an experiment at the crossroad between the experimental designs of Hennig-Schmidt et al. (2011) and Li et al. (2017). More precisely, we elicit altruism preferences of future physicians using a medical frame and standard experimental economics methods similar to, for example, Andreoni and Miller (2002), Fisman et al. (2007), Choi

⁴⁹The American Life Panel is a probability-based panel which is representative for the US-population and can be used to regularly interview participants over the internet.

et al. (2007), and Bruhin et al. (2019). Instead of relying on out-of-pocket expenses or monetary gains, our experimental design involves a series of medically framed choice tasks in which future physicians decide for 30 stylized patients on the treatment options rendering a profit for the physician and a health benefit for the patient. It is an important feature of our experimental decision situation that the receiving person being an actual patient is made salient. In particular, the patient health benefit is measured in monetary terms in the experiment, and real patients outside the experiment benefit from the medical students' decisions as the money was earmarked for surgery of cataract patients. This procedure on average elicited 25 percent of the per-patient cost of surgery from each participant. Our sample consists of 733 medical students from the University of Cologne, a major German public university teaching medicine. Students are spread according to the major progress steps of the six years of medical education in Germany: from freshmen, to pre-clinical studies (first and second year), clinical studies (third to fifth year), and practical year as assistant clinician in hospitals (sixth year).

Following the bulk of experimental elicitation of altruism preferences, we estimated a decision model based on a CES utility function in the fashion of Andreoni and Miller (2002) with two parameters, one capturing the altruism tradeoff and the other the equality-efficiency tradeoff. The size of the parameters estimated from the proposed sequence of binary choices informs us about the relative importance of different preference components, controlling for study progress. To account for medical students' characteristics which could potentially confound our estimates, we survey individuals about their socio-demographics, social and economic preferences according to Falk et al. (2016), personality traits (Gosling et al., 2003; Rammstedt and John, 2007; Ashton and Lee, 2009), and stated occupational preferences with an extensive post-experimental questionnaire.⁵⁰

Our structural estimation results show that, on the aggregate, medical students are considerably altruistic. Nevertheless, medical education significantly affects medical students' patient-regarding altruism. We find a non-linear, U-shaped relationship between their patient-regarding altruism and their progress in medical education. Compared to freshmen,

⁵⁰This paper is part of a broader panel study with medical students, who participate in the study up to four times in the course of their medical education.

medical students in the pre-clinical phase are more profit-oriented with the maximum profit orientation being observed during the clinical phase. Patient-regarding altruism slightly increases again in the practical year. The effects of study progress remain stable when controlling for medical students' gender, general altruism, other social and economic preferences, personality traits, and unobserved heterogeneity. Compared to a control group of non-medical students, medical students' patient-regarding altruism is significantly higher. To assess the predictive power of our preference estimates, we further link medical students' patient-regarding altruism to their future income expectations and their stated specialty choices. Lower patient-regarding altruism relates to higher estimates of their own future income expectations. We also find a stable link to specializing in pediatrics and surgery such that medical students opting for these specialties reveal significantly higher patient-regarding altruism.

The remainder of the paper is laid out as follows. Section 4.2 provides some background on medical education in Germany and on our sample. In Section 4.3, we present our experimental design and procedure. Section 4.4 describes our econometric strategy for estimating the behavioral model's parameters at different levels of aggregation. Section 4.5 presents the behavioral results and the structural estimation results. Finally, Section 4.6 concludes.

4.2 Background and sample

4.2.1 Medical education in Germany

The vast majority of prospective physicians in Germany is educated at one of the 36 public medical schools (e.g., Zavlin et al., 2017). The admission to medical education is centralized nationally by the non-profit governmental Foundation for Admission to Higher Education (*Stiftung für Hochschulzulassung*), and is highly competitive, as only about one out of five applicants is admitted to a German medical school.⁵¹ Admission criteria to medical schools, typically, are grades according to the General Certificate of Education (GCE), A-levels, waiting terms, and the applicants' performance in entry tests for studying medicine (TMS,

⁵¹For example, in the winter term 2018/2019, according to the Stiftung für Hochschulzulassung 43,631 prospective students applied to study medicine in Germany while only 9,232 places were available.

Test für Medizinische Studiengänge). At the time of data collection, 20% of the available places at medical school were assigned to applicants with the best GCEs and to applicants based on accumulated waiting terms, respectively. The remaining places (60%) were allocated based on a medical school's individual selection criteria (e.g., TMS).⁵²

The medical education in Germany is highly regulated. Structure, curriculum, and examination guidelines are standardized in the Medical Licensure Act (*Approbationsordnung für Ärzte*, ÄApprO, 2002) to ensure that all medical students obtain an appropriate and equivalent medical education.⁵³ Medical education lasts for at least six years and three months and concludes with the “Approbation”, the official German license to practice as a physician, upon successfully passing the physician state exam (*Staatsexamen*). Along the different parts of the physician exam, medical education in Germany typically comprises three phases: (i) pre-clinical phase, (ii) clinical phase, and (iii) practical year.

In the first two years of the medical studies, the pre-clinical phase, students are taught the basics of medicine and natural sciences and take part in a nursing internship. Traditionally, the pre-clinical phase concludes with the first part of the physician exam. Instead of one final exam at the end, medical students in Cologne take separate tests at different times of the pre-clinical phase which serve as an equivalence to the first part of the state exam. The subsequent clinical phase comprises a minimum of three years. In this more practical phase all relevant clinical subjects are covered and students gain first experiences in practicing medicine as physician-interns in hospitals and outpatient settings prior to taking the second part of the physician exam. Medical education concludes with the practical year, the aim of which is to familiarize students with practical work in clinics. The students spend four months each at the hospital's department of internal medicine, the department of surgery, and an elective department different from the former ones. After the practical year and having successfully completed the third part of the physician exam, medical students receive their license to practice medicine, and may start their actual specialization for a specific

⁵²In line with general guidelines implying a high weighting of GCE, every medical school can decide on applying their own selection criteria. At the University of Cologne, the internal selection is performed based on GCE (51%) and the applicants' performance in TMS (49%).

⁵³Note that slight and predefined deviations from the standardized course of study are possible (§41 ÄApprO) due to the medical education at the University of Cologne being accredited as a so-called model course of study.

field in medicine.

4.2.2 Our medical student sample

A total of 733 medical students of the University of Cologne participated in our study from April 2017 to December 2020. The sample consists of four cohorts: Freshmen in the first week of their medical studies who did not get any prior medical education at the University of Cologne, and students from each of the three study phases (pre-clinical, clinical, and practical year). Table 4.1 provides an overview on the composition of our sample: freshmen 35.3%, pre-clinical 32.1%, clinical 21.6%, and practical year, 11.1%. 74% of the observations were collected in 2017 (summer and winter term) with the average response rate being 15%.⁵⁴

Our sample consists of 440 (60%) females, the overall average age when starting

Table 4.1: Sample composition by study progress

	Freshmen	Pre-clinical phase	Clinical phase	Practical year
Curriculum	First week of medical studies	Basic science, nursing internship	Clinical topics, internship as a physician	Practical work in hospital
Year(s) of studies	0	1-2	3-5	6
Medical student sample ($N=733$)	259 35.3%	235 32.1%	158 21.6%	81 11.1%
Control group of non-medical students ($N=145$)	40 27.6%	23 15.9%	56 38.6%	26 ^a 17.9%

Notes. ^a6 years of studies and more

medical education was 20.7, and the share of Germans by nationality is 92.5%. The sample composition is a rather good approximation of the medical student populations in Germany and at the University of Cologne in terms of gender, age, nationality, and admission quotas, see Table 4.2.

In addition to our sample of medical students, we study a control group of 145 non-medical students of different majors such as business administration, economics, politics, law, history, linguistics, literature, pedagogy, and natural sciences enrolled at the University

⁵⁴For freshmen, pre-clinical, clinical, and practical year students response rates were 43%, 23%, 11%, and 4%, respectively, of those who were invited to participate in the study. We approached students at specific study terms, and, therefore, calculated the response rates based on the number of medical students in the respective study terms.

Table 4.2: Our medical student sample in context

	Our total sample	Comparison for 2017		
		Germany ^a	University of Cologne ^b	Our sample ^c
Female (%)	60.0	61.5	61.7	61.0
Age at starting medical education ^d	20.7	19.5 ^e	22.5 ^f	21.2
Share of Germans (%)	92.5 ^g	87.3	86.5 ^h	92.6 ⁱ
Admission quota ^j				
School-leaving grade (%)	21.8	20.0	20.0	21.9
Accumulated waiting terms (%)	9.9	20.0	20.0	10.3
University-specific selection criteria (%)	68.3	60.0	60.0	67.9

Notes. For the German student population descriptive statistics are only available for the winter term 2017/2018. Our sample comprises data from both the summer term and the winter term 2017 as the University of Cologne (UoC) is one of the few medical schools in Germany where students can enroll in both the summer and the winter term, and collecting our data started in April 2017. ^aGerman Federal Statistical Office (Statistisches Bundesamt, 2021). Data for the winter term 2017/2018: $n = 93,946$; ^bSummer term 2017 and winter term 2017/2018: $n = 6,034$; ^cSummer term 2017 and winter term 2017/2018: $n = 554$; ^dFreshmen only; ^eData only available for average age of graduates (Statistisches Bundesamt, 2018a). We, therefore, approximate the age at start of medical studies for overall Germany by subtracting the average study duration from the average age of graduates; ^fData available only for summer term 2017: $n = 3,000$; ^gDue to missing data: $n = 657$; ^hDue to missing data: $n = 539$; ⁱDue to missing data: $n = 539$; ^j The calculation of the admission quota rests on lower numbers than reported in Table 4.1, namely $n = 616$ for our total sample and $n = 507$ for the 2017 sample. Further deviations are due to procedural requirements in Germany, as some quotas are deducted from the total number of available German medical study places before allocating them to the applicants.

of Cologne. The cohorts match the medical student sample regarding the study progress in years, see Table 4.1.

4.3 The experiment

4.3.1 General design and decision situation

We introduce a novel experimental task in a stylized medical frame to elicit patient-regarding altruism. $N = 733$ medical students each decide in the role of a physician (i) and face $J = 2$ treatment alternatives (referred to as “A” and “B” in the instructions) for 30 stylized “patients” ($T = 30$ choice occasions). Physician own profit (payment to self) is represented by s_{jt} , and o_{jt} represents the patient benefit (payment to other) for treatment alternative j and patient t . Henceforth, we use the labels “physician” and “patient” to indicate the roles in our experiment.

Physician profit as well as patient benefit are expressed in monetary terms. While all subjects in the experiment make decisions in the role of physicians for stylized patients, real patients’ health outside the experimental setting is affected by their choices. Similar to earlier controlled experiments on physician behavior, the monetary equivalent of the

patient’s benefit resulting from the treatment alternative chosen is transferred to a charity and is earmarked for surgical treatment of cataract patients.⁵⁵ The treatment of a cataract patient costs about EUR 30. For procedural details, see Section 4.3.2.

Each of the 30 choice occasions implies a systematically varied trade-off between physician profit and patient benefit such that one treatment alternative is always more patient-regarding, see Table 4.3. The values for physician profit and patient benefit can take five values: EUR 3, 6, 9, 12, and 15.⁵⁶ The computerized experiment was programmed in ILIAS, a free software used as online learning platform in German universities. Medical students in Cologne are familiar with ILIAS, as it is commonly used for surveys and tests. The 30 choice occasions were shown in a pre-determined randomized order on subjects’ computer screens. A subject’s total payoff consisted of a physician profit (from a randomly selected patient) and a lump-sum payment for filling in the post-experimental questionnaire (EUR 5).

4.3.2 Experimental protocol

The recruitment procedure was as follows. Sessions with freshmen were conducted during the first (introductory) week of their studies prior to starting their medical education. Besides freshmen, we approached pre-clinical students mainly at the end of their first year or in their second year, clinical students mainly in their fourth year as well as practical-year students in their sixth year.

In total, we conducted 16 experimental sessions between April 2017 and December 2019. We run 11 laboratory sessions in a large lecture hall equipped with computer terminals at the medical school of the University of Cologne. The remaining five sessions were conducted online for a period of 10 to 26 days in order to reach students across all study phases who

⁵⁵This procedure was introduced by Hennig-Schmidt et al. (2011) and has been applied in several experiments in health economics, as it embeds an incentive for subjects in the lab to account for real patients’ health outside the lab. Equivalent mechanisms have been employed in recent behavioral experiments in health analyzing physician behavior (Hennig-Schmidt et al., 2011; Hennig-Schmidt and Wiesen, 2014; Godager et al., 2016; Brosig-Koch et al., 2016, 2017a, 2020; Byambadalai et al., 2019; Di Guida et al., 2019; Martinsson and Persson, 2019; Huesmann et al., 2020; Wang et al., 2020; Waibel and Wiesen, 2021). In Kesternich et al. (2015) and Lagarde and Blaauw (2017), subjects could choose from several (medical) charities to which a donation should be transferred.

⁵⁶The specific values of the treatment alternatives were chosen to guarantee that participants’ average earnings correspond to the hourly wage of a student assistant at the University of Cologne (EUR 10). We excluded values of zero to avoid end points. We used the command ‘dcreate’ in STATA 14.0 (Hole, 2015) to guide the parameterization of our choice occasions.

Table 4.3: Physician profit and patient benefit for treatment alternatives A and B for the 30 patients

Patient t	Treatment A		Treatment B	
	Profit s_{At}	Benefit o_{At}	Profit s_{Bt}	Benefit o_{Bt}
1	3	15	6	9
2	3	15	9	9
3	3	15	15	3
4	3	15	6	6
5	9	15	12	12
6	6	9	15	3
7	15	3	6	9
8	3	15	6	3
9	3	15	12	6
10	9	9	3	15
11	3	9	9	3
12	15	3	3	15
13	3	15	12	12
14	3	12	12	3
15	6	12	9	6
16	3	9	6	6
17	12	12	15	9
18	3	12	15	3
19	9	6	3	12
20	6	6	3	15
21	12	12	3	15
22	12	3	3	9
23	15	6	6	12
24	6	3	3	6
25	3	9	15	3
26	6	9	3	15
27	6	9	9	6
28	15	6	9	12
29	15	9	9	15
30	6	12	15	3

were not able to participate in the laboratory sessions. We collected 457 (62.4%) observations via laboratory sessions and 276 (37.6%) observations via online sessions. Between November 2019 and January 2020, we run an online experiment with a control group of 145 non-medical students of the University of Cologne, who were recruited via the online recruiting system ORSEE (Greiner, 2015).

Prior to the experiment, subjects received detailed information on the data protection, the experimental decision task, the procedure and the payment process. For more detail, see the instructions provided in Appendix D.1.1. All subjects decided for the same 30 stylized patients. After subjects had taken their decisions, they were asked to complete a

comprehensive questionnaire (see Section 4.3.3)⁵⁷.

It took subjects, on average, about 60 minutes to complete the decision tasks and the questionnaire. On average, medical students earned EUR 12.11 (profit EUR 7.11 plus EUR 5 for completing the questionnaire), and non-medical students were paid EUR 11.68 (profit EUR 7.68 plus EUR 4). The average patient benefit amounted to EUR 7.89 for medical and EUR 7.32 for non-medical students. In total, EUR 6,846 were transferred to the *Christoffel Blindenmission*, a charity that used the money exclusively for financing cataract surgery by their own ophthalmologist staff in developing countries. Our study, thus, enabled the treatment of 228 cataract patients at the average cost for a surgery of EUR 30. As subjects' decisions realize an average patient benefit of EUR 7.80, this makes up for one fourth of the total cost of an eyesight-restoring surgery by each participant.

4.3.3 Post-experimental questionnaire

A comprehensive post-experimental questionnaire provided data that allow us to analyze how medical students' characteristics relate to their patient-regarding altruistic behavior. In addition to students' main characteristics (study cohort, gender, and age; recall Section 4.2.2), we collect information on subject's personality traits, on their social and economic preferences, and on their future work-related preferences (e.g., preferred specialty and future income expectations).⁵⁸

To get a comprehensive picture on subjects' preferences, we elicited social and economic preferences through experimentally validated survey-based methods according to Falk et al. (2016, 2018). Social preferences comprise preferences for general altruism, trust, positive and negative reciprocity, and economic preferences comprising time and risk preferences. Note that the measure for general altruism is of particular relevance for our study as it enables us to relate subjects' general altruism to subjects' incentivized patient-regarding altruistic choices. Additionally, we elicited subject's personality traits extraversion, agreeableness, conscientiousness, neuroticism/emotionality, and openness using the 11-item short-version of

⁵⁷All questionnaire items, which were only applicable for medical students, were dropped for the non-medical student sample.

⁵⁸As our study is part of a broader panel study, the post-experimental questionnaire comprises several more items. In this paper, we only mention those questionnaire items relevant for the purpose of this study.

the Big Five Inventory (Gosling et al., 2003; Rammstedt and John, 2007). From winter term 2018 onwards, we use the more detailed 60-item HEXACO Personality Inventory (Ashton and Lee, 2009).⁵⁹ For a detailed description of the questionnaire items, see Table D.1.1 in Appendix D.1.2.

4.4 Empirical strategy

4.4.1 Behavioral model of altruism

To characterize the distribution of altruism at the aggregate, we employ a behavioral model of altruism following previous literature (e.g., Andreoni and Miller, 2002; Fisman et al., 2007; Bruhin et al., 2019). As a utility function u_i for subject i we consider a constant elasticity of substitution (CES) parametric form, defined as:

$$u_i(s, o, a_i, r_i) = (a_i s^{r_i} + (1 - a_i) o^{r_i})^{\frac{1}{r_i}}, \quad (4.1)$$

where $a_i \in [0, 1]$ represents the weight a physician puts on her own profit, correspondingly $1 - a_i$ represents the weight a physician puts on the patient's health benefit, and $r_i < 1$ reflects the elasticity of substitution between own profit and the patient's health benefit. The elasticity of substitution is defined as $\sigma_i = \frac{1}{r_i - 1}$.

Depending on the values of a and r , subjects belong to different preference types. For example, a subject with a equal to one is a purely selfish type, because she does not put any weight on the patient's health benefit. (i) When $a \in (0.5, 1]$, the subject places more weight on own profit compared to the patient health benefit; (ii) when $a = 0.5$, the individual places the same weight on own profit and the patient's benefit; (iii) when $a \in [0, 0.5)$, the individual places less weight on own profit compared to the patient's benefit.

The parameter r reflects the curvature of the altruistic indifference curves. It ranges from $-\infty$ to 1: (i) when $r \in (0, 1]$, preference is weighted toward increasing the sum of own profit and patient benefit, with CES approaching a perfect-substitute utility function as

⁵⁹The HEXACO Personality Inventory elicits the same personality traits as the 11-item short-version Big Five Inventory, yet with 10 items each. As the additional trait honesty-humility is not included in the Big Five Inventory and, therefore, data is limited to 179 medical students, we neglect this trait in our subsequent analyses.

$r \rightarrow 1$; (ii) when $r \in (-\infty, 0)$, the individual's distributional preference is instead weighted toward reducing the difference in own profit and patient benefit with CES approaching the Rawlsian or Leontief utility function $\min\{as, (1-a)o\}$ as $r \rightarrow -\infty$; (iii) when $r \rightarrow 0$, CES is approaching the Cobb-Douglas function $s^a o^{(1-a)}$, in which case the allocation to patient benefit o and physician s is constant.

4.4.2 Structural estimation

In this section, we describe our strategy for estimating the parameters of the behavioral model. To estimate the parameters of this model, a and r , we apply McFadden's (1981) random utility model for discrete choices. In the experiment, N subjects faced $J = 2$ alternatives on $T = 30$ choice occasions (patients). We assume that subjects choose the alternative that maximizes their utility for each choice occasion. The random utility of subject i from choosing alternative j in a choice occasion (for a patient) t is defined as:

$$U_{ijt} = u_i(s_{jt}, o_{jt}) + \epsilon_{ijt}, \quad (4.2)$$

where $u_i(s_{jt}, o_{jt})$ is the deterministic utility of the allocation of own profit to the physician s_{jt} and the benefit to the patient o_{jt} and the error term is distributed extreme value. The choice probability between option $j = 1$ and option $j = 2$ for subject i in choice occasion t , can be written:

$$P_{it}(a_i, r_i, \mu_i) = \frac{1}{1 + e^{\frac{u_i(s_{1t}, o_{1t}, a_i, r_i) - u_i(s_{2t}, o_{2t}, a_i, r_i)}{\mu_i}}}.$$

Where $\mu_i > 0$ is a noise parameter with the following interpretation: when $\mu_i \rightarrow 0$ the choice becomes deterministic, when $\mu_i \rightarrow +\infty$ the choice becomes entirely random.

For the sake of simplicity, we denote $\theta_i = (a_i, r_i, \mu_i)$ the set of parameters for the CES preference functional. Let $y_{it} = 1$ if the subject chooses option $j = 2$ for choice t (and $y_{it} = 0$ otherwise), the likelihood associated with this choice is:

$$P_{it}(\theta_i)^{y_{it}}(1 - P_{it}(\theta_i))^{1-y_{it}}$$

and the likelihood associated with a choice sequence T for subject i with parameters θ_i, μ_i is written as:

$$P_i(\theta_i) = \prod_{t=1}^T P_{it}(\theta_i)^{y_{it}}(1 - P_{it}(\theta_i))^{1-y_{it}}.$$

We use a series of transformation functions to account for theoretical restrictions on parameters. First, the estimation procedures account for parameter constraints on noise with an exponential transformation ($\mu_i = g^\mu(\zeta_i^\mu) = \exp(\zeta_i^\mu)$) of the parameter value to guarantee the noise parameter is positive. For CES, the estimation procedures account for parameter constraints with transformation of the estimated parameter values ζ_i^a and ζ_i^r . For parameter a_i we use a logistic function ($a_i = g^a(\zeta_i^a) = \frac{1}{1+\exp(-\zeta_i^a)}$) guaranteeing a_i , which is a share, is always between zero and one. For parameter r_i we use a translated negative exponential function ($r_i = g^r(\zeta_i^r) = 1 - \exp(-\zeta_i^r)$), guaranteeing r_i is always lower than one. In vector notation, for the CES preference functional, $\theta_i = g(\zeta_i)$, with $\zeta_i = (\zeta_i^a, \zeta_i^r, \zeta_i^\mu)$ and $g() = (g^a(), g^r(), g^\mu())$.

The regression tables in the main text (see Section 4.5) report the value of the transformed regression coefficients in preference parameters and noise, i.e. when regression coefficients are transformed back to the original scale. In other words, the constant is defined by $\theta_i = g(\zeta_i)$ and represents median parameters.

The first version of the random utility model pools the data and estimates aggregate parameters, a and r , that are representative for all subjects. These aggregate estimates represent the most parsimonious characterization of altruism.

For aggregate estimations, all individual share the same parameters: $\theta_i = \theta, \forall i$,

$$LL(\theta) = \sum_i \ln(P_i(\theta)).$$

The estimation procedure maximizes the log-likelihood with respect to unconstrained pa-

rameters. The (grand) log-likelihood to be maximized with respect to ζ is:

$$LL(\zeta) = \sum_i \ln(P_i(g(\zeta))).$$

Reported standard errors are clustered at the individual level and computed with the (multivariate) Delta method (see Appendix D.2 for details).

Aggregate estimations provide a parsimonious characterization of preference but neglect heterogeneity. In order to account for observed heterogeneity between participants we include in the structural estimation several covariates relating to study cohort, gender, preferences under risk, time preferences, social preferences and a series of personality measures. This defines six econometric models, each associated with a different set of covariates: Model (1) is the base model without covariates, Model (2) adds term cohorts (our main research interest) as explanatory variable, Model (3) adds gender as explanatory variable, Model (4) adds the measure of the questionnaire item 'general altruism' as explanatory variable, Model (5) adds measures of risk aversion, discounting, trust, positive reciprocity and negative reciprocity as explanatory variables, and Model (6) adds personality measures as explanatory variables.

More formally, let X_i denote the column vector of K regressors for individual i . The first element of X_i is a one. We denote η by the $3 \times (K + 1)$ matrix of parameters such that:

$$\zeta_i = \eta X_i.$$

The (grand) log-likelihood to be maximized with respect to η is:

$$LL(\eta) = \sum_i \ln(P_i(g(\eta X_i))).$$

The regression tables in the main text report the value of the preference and noise

parameters when regression coefficients are transformed back to their original scale.⁶⁰

4.5 Results

4.5.1 Descriptives and non-parametric analyses

To start with our analysis, we first focus on descriptives and present non-parametric analyses. Table 4.4 shows the summary statistics for the proportion of patient-regarding choices (*prcs*, Panel A). The *prc* refers to the treatment alternative which provides the patient with the higher health benefit in each choice occasion. We also report summary statistics for subjects' characteristics (Panel B). Medical students decide in a patient-regarding way by, on average, taking 56.3% *prcs*. When differentiating between study progress, we find a U-shaped relationship. Freshmen are the most patient-regarding cohort (66.0%). Patient orientation decreases in the pre-clinical phase (53.7%), reaches its minimum in the clinical phase (46.0%), and rises to 52.0% in the practical year.

Patient-regarding behavior of freshmen is significantly higher compared to the three other cohorts ($p < 0.001$, *t*-test). *prcs* in the pre-clinical phase are significantly higher than in the clinical phase ($p = 0.009$) but do not differ significantly between practical-year students and students in the pre-clinical or clinical phase ($p > 0.170$).

Figure 4.1 illustrates the distributions of *prcs* by study cohorts. Except for comparing clinical-phase and practical-year students, the Kolmogorov-Smirnoff test rejects the hypothesis of identical distributions for all cohort comparisons ($p < 0.05$). Figure 4.1 also shows how medical students differ in their behavior. Pure profit-maximizers, who do not make any *prc*, are located at the bottom of the graphs. For freshmen, their share is lowest (2.7%). It amounts to 3.0% in the pre-clinical phase, but increases to 15.8% in the clinical phase and to 13.6% in the practical year. On the other hand, the share of pure altruists, those always

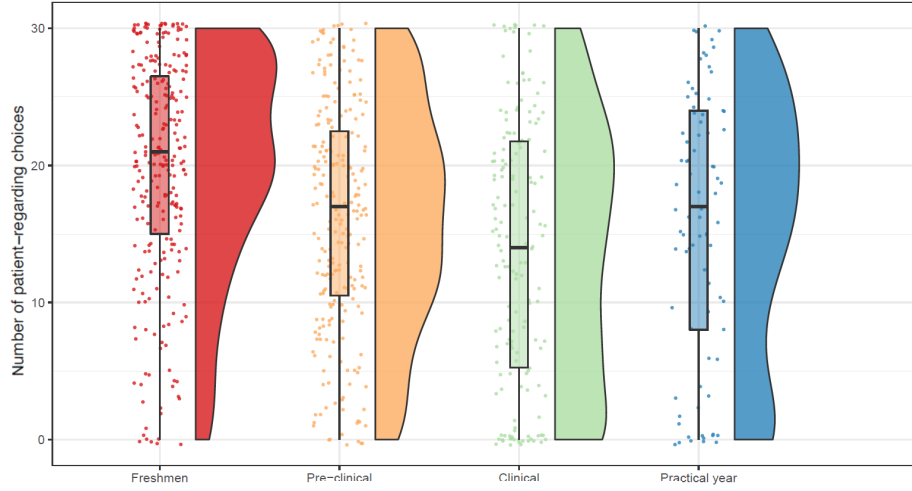
⁶⁰In other words, the constants are defined by $\theta_0 = g(\eta_0)$ and represents median parameters. For the dummy variables, such as gender (or cohort), the tables with marginal effects report $g(\eta_0 + \eta_{female}) - g(\eta_0)$, the (partial) effect setting the female dummy to one. The same applies to cohort. For preference parameters, distributed with a range of 1 (from -0.5 to 0.5 for risk aversion, with 0 indicating risk neutrality, from 1 to 0 for discounting, with 1 indicating no discounting and from 0 to 1 for other preference items), marginal effects are equal to $g(\eta_0 + \eta_{\text{preference item}} \times 0.1) - g(\eta_0)$. For personality measures, distributed over the support $[-1, 1]$, marginal effects are equal to $g(\eta_0 + \eta_{\text{personality item}} \times 0.2) - g(\eta_0)$. Standard errors are computed with the Delta method.

Table 4.4: Descriptive statistics of medical students' behavior and characteristics

	Mean M	s.d.	N
A. Patient-regarding choices			
Total sample	16.9 (56.3%)	9.0	733
Freshmen	19.8 (66.0%)	8.1	259
Pre-clinical	16.1 (53.7%)	8.3	235
Clinical	13.8 (46.0%)	9.6	158
Practical Year	15.6 (52.0%)	9.6	81
B. Characteristics			
<i>Social and economic preferences</i>			
Altruism	0.38	0.17	729
Trust	0.57	0.24	729
Positive reciprocity	0.36	0.18	729
Negative reciprocity	0.47	0.16	729
Risk aversion	0.07	0.15	731
Time discounting	0.27	0.16	731
<i>Personality traits</i>			
Agreeableness	0.09	0.37	729
Conscientiousness	0.39	0.35	729
Extraversion	0.25	0.41	729
Neuroticism/emotionality	-0.08	0.43	729
Openness	0.27	0.44	729

Notes. This table presents summary statistics on the number of patient-regarding choices and on subject's characteristics, the latter comprising social and economic preferences according to Falk et al. (2016; 2018), personality traits by the Big Five Inventory, (Gosling et al., 2003; Rammstedt and John, 2007) or the HEXACO Personality Inventory (Ashton and Lee, 2009). Altruism, trust, positive and negative reciprocity are measured on a $[0, 1]$ -scale with 0 being the theoretical minimum and 1 the theoretical maximum. Risk aversion is transformed such that 0 implies risk neutrality, a positive value entails risk aversion and a negative value risk seeking. Time discounting being 0 entails patience, while a positive value implies impatience. All personality traits are measured on a $[-1, 1]$ -scale. See Table D.1.1 in Appendix D.1.2 for a detailed description of all variables. The lower number of observations in Panel B is due to subjects leaving the survey before completing the questionnaire.

Figure 4.1: Distributions of patient-regarding choices by cohorts



Notes. This figure shows distributions and box plots for number of patient-regarding choices by study cohort. Pure profit-maximizers are located at the bottom of the graph and pure altruists at the top.

choosing the high-benefit alternative, is highest for freshmen (12.4%), while it is 4.3% in the pre-clinical phase, 7.0% in the clinical phase and 4.9% for practical year students.

So far, our analysis provides evidence that the majority of medical students reveal preferences, which deviate from pure profit-maximization, and that patient benefit plays an important role in their treatment decisions. By contrast, our control group of non-medical students is significantly less patient-regarding in all study progress stages. For further details on the choice behavior and descriptive statistics of our control group, see Appendix D.3.

We now turn to characteristics of the individual participants that were elicited in the questionnaire part of our study (see Section 4.3.3). Panel B in Table 4.4 shows the descriptive statistics on subjects' social and economic preferences (general altruism, trust, positive and negative reciprocity, risk, and time discounting) as well as on subjects' personality traits (agreeableness, conscientiousness, extraversion, neuroticism/emotionality, and openness).

Altruism, trust, positive and negative reciprocity are measured on a $[0, 1]$ -scale with 0 being the theoretical minimum and 1 the theoretical maximum. Risk aversion is transformed such that 0 implies risk neutrality, a positive value entails risk aversion and a negative value risk seeking. Time discounting being 0 entails patience, while a positive value implies impatience.

For our medical student sample, the general altruism measure is $M_{altruism}=0.38$, which is below the theoretical midpoint of 0.50 and indicates that on the stated preference level, the students tend to be more selfish than altruistic. Our sample tends to be trusting ($M_{trust} = 0.57$), and is slightly more positively reciprocal than being negatively reciprocal ($M_{pos.recipr.}=0.47$, $M_{neg.recipr.}=0.36$). $M_{risk}=0.07$ points to a risk aversion of our participants, while the positive value for time discounting ($M_{timedisc.}=0.26$) indicates impatience.

All personality traits are measured on a $[-1, 1]$ -scale. Regarding agreeableness and neuroticism/emotionality, the sample means are close to the neutral midpoint ($M_{agr.}=0.09$ and $M_{neurot.}=-0.08$). The positive values for the remaining personality traits reveal that our students are rather conscientious, extraverted, and open ($M_{conscient.}=0.39$, $M_{extrav.}=0.25$ and $M_{openn.}=0.27$). The impact of preferences and personality traits on subjects' patient-regarding choices will be accounted for in the structural estimation analyses.

4.5.2 Structural estimation with observed heterogeneity

We next present the aggregate estimation results. Table 4.5 shows the estimation results with transformation of the estimated parameters into preference parameters, noise parameters, and marginal effects.⁶¹ In our base model without covariates, a indicates a rather moderate profit orientation. Medical students put a weight of about one third on their own profit and two thirds on the patient health benefit. Parameter r is negative, implying that medical students express a tendency for inequity aversion; see Model (1) of Table 4.5.

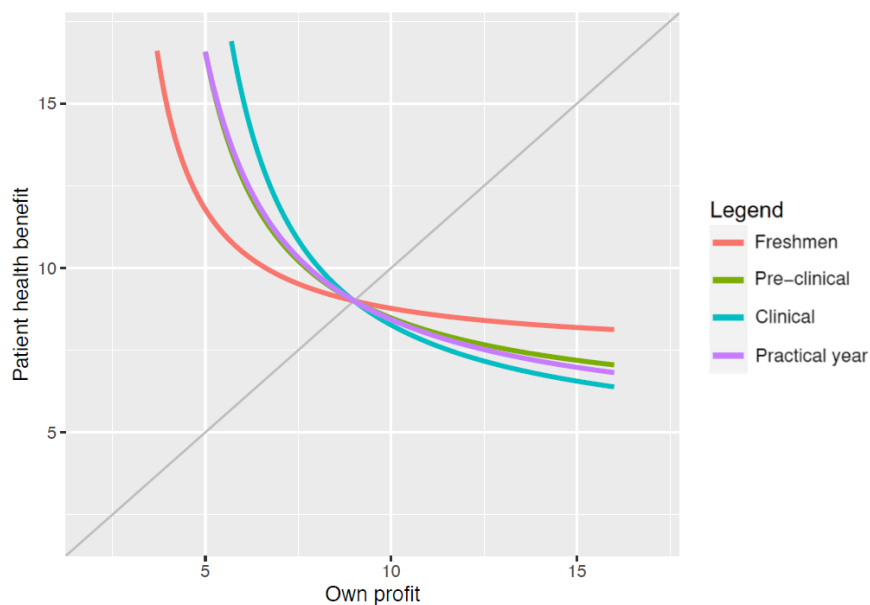
Our estimation results support our non-parametric analyses. Medical students' patient-regarding altruism significantly declines with progress in medical education. Recall that a characterizes participants' own-profit orientation, and, thus, a negative (positive) coefficient for a implies an increase (decrease) in the weight subjects put on the patient benefit. Model (2) of Table 4.5 shows, that compared to freshmen (our reference category), the medical students in the pre-clinical phase are more profit-oriented and profit orientation is highest during the clinical study phase. Only in their practical year, medical students' patient-regarding altruism slightly increases again compared to the clinical studies. Figure 4.2 shows

⁶¹The estimation results of the aggregate model without transformation of the estimated parameters are shown in Table D.4.1 of Appendix D.4.1.

the typical indifference curves for the different study cohorts based on parameter estimates from Model (2). This effect of the study phase remains stable when controlling for medical students' gender, general altruism, other social and economic preferences, and personality traits; see Models (2) to (5) of Table 4.5. For an illustration of the heterogeneity implied by the different set of covariates, see Figure D.4.1 in Appendix D.4.1.

Our estimations show that female medical students are more altruistic on behalf of their patients than male medical students; see Model (3) of Table 4.5. We also observe that patient-regarding altruism is positively related to Falk et al.'s general altruism measure meaning that medical students with higher general altruism put significantly less weight on their own profit compared to the patient's health benefit. These findings are also robust when controlling for other social and economic preferences, as well as personality traits; see Models (4) to (6) of Table 4.5.

Figure 4.2: Indifference curves for different study cohorts



Notes. This figure shows the indifference curves between own profit and patient health benefit for the different study cohorts based on CES preference parameter estimates from Model (2) of Table 4.5.

Overall, medical students reveal inequity averse preferences, as indicated by the estimates for the parameter r ; see Model (1) of Table 4.5. Also, estimates for r tend to increase

with medical students' term but with no firm results when adding further controls to the regressions. Further, estimations show that women and individuals with higher general altruism have lower values of r ; see Models (2) to (6) of Table 4.5. Noise μ tends to be lower for pre-clinical students and larger for students in practical years.

Table 4.5: Aggregate estimations, preference parameters, noise and marginal effects, CES preferences

Model:	(1)	(2)	(3)	(4)	(5)	(6)
a						
Constant	0.339*** (0.016)	0.209*** (0.024)	0.293*** (0.035)	0.531* (0.052)	0.708*** (0.090)	0.739*** (0.109)
Pre-clinical		0.168*** (0.035)	0.163*** (0.041)	0.154*** (0.048)	0.095*** (0.047)	0.095*** (0.051)
Clinical		0.258*** (0.039)	0.257*** (0.043)	0.213*** (0.047)	0.122*** (0.050)	0.124*** (0.053)
Practical year		0.189*** (0.056)	0.138*** (0.084)	0.163*** (0.070)	0.075*** (0.068)	0.081*** (0.064)
Female			-0.111*** (0.023)	-0.064*** (0.031)	-0.041*** (0.027)	-0.044*** (0.040)
General altruism				-0.079*** (0.011)	-0.056*** (0.013)	-0.054*** (0.018)
r						
Constant	-0.956*** (0.097)	-1.240*** (0.137)	-0.493*** (0.168)	-0.354*** (0.197)	-0.485*** (0.423)	-0.364*** (0.489)
Pre-clinical		0.342* (0.218)	0.149* (0.189)	0.132** (0.167)	0.181*** (0.218)	0.206*** (0.232)
Clinical		0.393* (0.273)	0.183 (0.240)	0.254*** (0.188)	0.276*** (0.245)	0.332*** (0.240)
Practical year		0.486* (0.376)	-0.092 (0.477)	0.100 (0.341)	-0.118 (0.545)	-0.014 (0.524)
Female			-1.066*** (0.205)	-0.490*** (0.186)	-0.489*** (0.243)	-0.424*** (0.272)
General altruism				-0.101*** (0.028)	-0.119*** (0.063)	-0.105*** (0.073)
μ						
Constant	2.623*** (0.092)	2.519*** (0.128)	2.313*** (0.170)	2.247*** (0.308)	2.007*** (0.487)	1.951*** (0.528)
Pre-clinical		-0.198 (0.194)	-0.171 (0.217)	-0.194** (0.224)	-0.149* (0.212)	-0.275*** (0.254)
Clinical		0.041 (0.265)	0.112 (0.307)	-0.106 (0.286)	-0.075 (0.266)	-0.158* (0.279)
Practical year		0.373 (0.363)	0.787*** (0.650)	0.552*** (0.572)	0.951*** (0.691)	0.768*** (0.741)
Female			0.127 (0.198)	-0.128 (0.200)	-0.064 (0.176)	-0.006 (0.215)
General altruism				0.041* (0.048)	0.026 (0.054)	0.017 (0.055)
Soc./econ. preferences	No	No	No	No	Yes	Yes
Personality traits	No	No	No	No	No	Yes
N	733	733	733	729	729	729
Log-likelihood	-13,331.09	-13,002.64	-12,884.71	-12,394.66	-12,028.09	-11,997.39

Notes. This table shows the estimation results of the aggregate model for the CES preference functional with study progress, gender, and general altruism in the set of covariates. Model (5) and (6) control for social and economic preferences and Model (6) for personality traits to account for observed heterogeneity. For the estimates of the full list of covariates, see Table D.4.2 in Appendix D.4.1. Standard errors are clustered at the individual level. Differences in the number of observations in Models (4-6) is due to missing data on some questionnaire items. *p<0.10; **p<0.05; ***p<0.01.

4.5.3 Robustness of results

Our next focus is on the robustness of our estimation results. In particular, we test whether results may be sensitive to the choices we have made on our econometric strategy. One potential source of sensitivity is how we extend our econometric model to account for individual heterogeneity. To check for this sensitivity, we consider several alternative econometric specifications. This section provides a brief overview on alternative econometric approaches to account for individual heterogeneity. Details of each of these econometric models and the respective estimation results can be found in Appendices D.4.2 to D.4.4.

First, an alternative direction to account for heterogeneity identifies distinct preference (and noise) types with a finite mixture model without covariates. We estimated finite mixture models with three types ($C = 2, 3, 4$). These finite mixture models apply an endogenous classification procedure of distinct preference types and show that the heterogeneity in preferences types is substantial; for methodological details and estimation results, see Appendix D.4.2. This direction supports our econometric approach to account for (observed) heterogeneity in the aggregate estimation as the results support discrimination between purely altruistic and moderately strong social preferences.

Another potential direction to account for heterogeneity between subjects incorporates observed and unobserved heterogeneity in a random coefficient model. We use Bayesian methods for our estimations. The estimation results on our a parameter of Table D.4.7 in Appendix D.4.3 are consistent with those reported in Table 4.5 based on the aggregate estimation. This suggests that our findings of selfish preferences to increase with study progress and with a maximum attained in the clinical phase are robust to estimating unobserved heterogeneity with a random coefficient model.⁶²

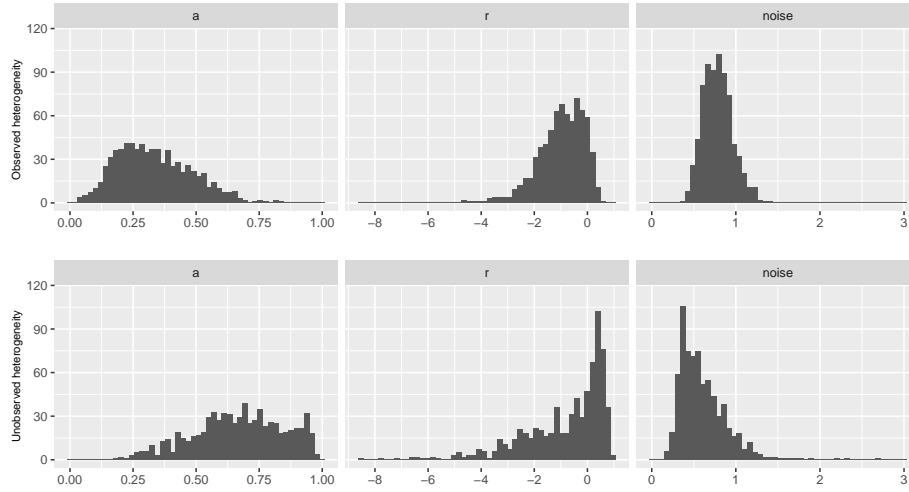
Figure 4.3 shows a visual comparison between the distributions of observed vs. unobserved heterogeneity for the preference parameters a and r and the noise parameter μ , when the full list of covariates is included.⁶³ The comparison of the two distributions

⁶²Tables with a full list of all estimated parameters are provided in Appendix D.4.3, Table D.4.5 and Table D.4.7.

⁶³Because histograms are based on non-linear transformations of estimated parameters and mean individual-level draws across iterations for observed vs. unobserved heterogeneity, there is no natural comparison benchmark between each pair of distributions for noise and preference parameters.

shows that observed characteristics generate a rich variation in preference parameters a and r .

Figure 4.3: Observed vs. unobserved heterogeneity in the random coefficient model



Notes. This figure shows the distributions of individual preference and noise parameter based on CES preference parameter estimates from Model (6) with the random coefficient model. The top row represents observed heterogeneity, the bottom represents unobserved heterogeneity. Unobserved heterogeneity is based on transformed mean individual-level draws across iterations for the underlying multivariate normal distribution with all covariates equal to zero.

A last alternative is to estimate the preference and noise parameters for each individual separately. The summary statistics confirm that, on average, concerns for patients are higher than for own profit. In particular, the mean of the individual a estimates is close to the aggregate estimate. Table D.4.8 provides an overview of descriptive statistics for the individual parameter and preference estimates.

Our results may be sensitive to our choice of the structural model of altruistic preferences, the CES preference utility function. To examine the robustness of our findings to this potential sensitivity, we consider a Fehr and Schmidt (1999) parametric form as another variant for the utility function. The social preference model by Fehr and Schmidt captures preferences of disadvantageous and advantageous inequality through two distinct parameters; for more details on the behavioral model, see Appendix D.5.1. Overall, applying a Fehr and Schmidt preference functional supports our main findings that patient-regarding altruistic

preferences decrease with study progress, with the lowest altruistic preferences in students in the clinical study phase. Within the framework of the utility function, we find that our medical students are extremely averse to advantageous inequality, i.e., averse of them as physicians to be ahead of the patient, with clinical phase students being the least averse. At the same time, our medical student sample is also somewhat (albeit less) averse to disadvantageous inequality, i.e., averse of being behind, with the highest aversion attained in clinical phase students. We provide results for the aggregate estimation in Appendix D.5.2 and for the random coefficient model in Appendix D.5.3.

Finally, we additionally perform all estimations including the control group of non-medical students. In Appendix D.4.5, we present all additional analyses for CES preferences in detail. The estimation results of each analysis are consistent with our presented main results. Nevertheless, we find that non-medical students behave less patient-regarding compared to medical students.

4.5.4 Patient-regarding altruism, income expectations, and specialty choices

Previous literature mainly focuses on the importance of income differences for choosing a specific specialty. While most studies report that a higher expected income positively affects this choice, the reported effect sizes are rather small (Bazzoli, 1985; McKay, 1990; Nicholson, 2002; Thornton and Esposto, 2003; Gagné and Léger, 2005). Moreover, a range of non-monetary factors, such as expected working hours (McKay, 1990), a regular working schedules and generous annual leave (Thornton and Esposto, 2003), a controllable lifestyle (Dorsey et al., 2003), intellectual content (Harris et al., 2005), and procedural work or academic opportunities (Sivey et al., 2012), are further reported to play an important role in the choice for a specialty.

Empirical evidence, however, is inconclusive on the kind of non-monetary factors that are most influential for selecting a particular specialty and on the impact of medical students' personality and individual preferences therein. In their literature review, Borges and Savickas (2002) summarize that evidence on a link between personality traits and specialty choices is rather weak, as there seems to be no specific specialty where all physicians show a unique pattern of personality traits.

To the best of our knowledge, Li (2018) is the first to link experimentally measured altruistic preferences to specialty choices of future physicians. She finds that lower altruism relates to selecting into high-income specialties. In order to do so, she constructs a binary measure grouping medical students' stated specialties into a low- or high-income group. Assigning income to specialties is based on two particular data sources. The cutoff for defining high-income specialties is having an annual average income of at least \$300,000. This is close to the median income of specialties chosen by the medical student subjects in the sample (\$304,000). This approach implies, however, that, first, the classification of income groups is not highly sensitive to the chosen source of income data and, second, that medical students are aware of the income they may be able to earn in the different specialties. It is not implausible to question that particularly those future physician with a lower emphasis on profit orientation might lack the particular knowledge about the income distribution across specialties. Also, it is questionable whether a medical students' choice for a specialty she would like to practice in should be narrowed down to a binary choice between a high- or a low-income group.

The assumptions Li (2018) made for classifying specialties to income groups based on average statistics data may hold for the US; they can be questioned, however, to hold for Germany. Here high income variations across specialties do exist in the outpatient care sector with differences in the net income for resident physicians up to EUR 212,000 (own computations based on Statistisches Bundesamt, 2018b, Tabelle 3.2, p. 21). This does, however, not apply for physicians employed in hospitals. Their salaries typically follow fixed payment schedules based on collective agreements. Thresholds for higher salary groups are based on physician's experience and length of affiliation to the respective hospital career (Marburger Bund, nd). Moreover, even when focusing solely on the average incomes of resident physicians, a classification into either a high- or low-income specialty group would appear arbitrary due to limited data availability for each specialty and divergent classifications of different data sources, like official statistics of Statistisches Bundesamt (2018b) or the "Physician Practice Panel" (Zentralinstitut für die Kassenärztliche Versorgung, 2016). Given our concerns, we follow an approach different to Li (2018).

We, first, investigate how the altruistic preferences of our medical student sample relate

to their individual future income expectations.⁶⁴ In particular, they had to indicate their expected monthly net income five years after having completed their specialty education and state the probability of falling into each of five income categories given a full-time job: (1) < EUR 3,000, (2) EUR 3,000 to 3,999, (3) EUR 4,000 to 4,999, (4) EUR 5,000 to 5,999, (5) > EUR 6,000. For each subject, we calculate an expected value for her future income expectations derived as the sum of the stated probabilities multiplied by the mean income of the respective category.⁶⁵ Second, we study how their altruistic preferences relate to their stated choice for a certain specialty without classifying specialties to any income groups.

Overall, our medical student sample ($N=693$) expects to earn on average EUR 4,427 net per month (s.d. 737). Hereafter, we split the continuous expected income variable at the median, dividing our sample into discrete groups to facilitate the interpretation of our estimation results. In particular, the dummy variable “Expected income” equals 1 in case a medical student expects a future income above the median (EUR 4,400), and 0 otherwise.

We add the income variable to our previous models (described in Section 4.5.2) in order to account for observed heterogeneity in expected future income when estimating our altruistic preferences. For medical students with future income expectations above the median, the estimated preference parameter a is significantly higher, see left column of Panel A and B of Table 4.6. As an increase in a implies a higher profit orientation, medical students who expect to earn relatively more in their future put significantly more weight on their own profit compared to the patient’s health benefit.

We finally study how our altruism measure relates to the stated speciality choices. Overall, the four most frequently stated specialties which were chosen by more than 10% of our medical student sample are surgery ($N= 137$, 19%), internal medicine ($N= 110$, 15%), pediatrics ($N= 97$, 13%), and neurology/psychiatry ($N= 84$, 12%). The remaining

⁶⁴Subject’s own specifications on their expected future income support our concerns to follow the approach of (Li, 2018). Comparing the own future income expectations of those students who state a preference for a specialty which would be classified into a high-income group based on Statistisches Bundesamt (2018b) to those stating a low-income group specialty reveals no consistent pattern. In particular within the high-income group, 54% expect to earn an income above the median in the future. The respective number is 51% in the low-income group.

⁶⁵In order to keep the range per category constant, we used EUR 2,500 and 6,500 as average values for the lower and upper bound. For the cases ($N= 82$) in which the sum of indicated probabilities doesn’t add up to 100 percent, we transformed the scaling according to the probability sum. We note that our results could be sensitive to the choices we have made in constructing the expected income variable.

Table 4.6: Aggregate estimations and random coefficient model, preference parameters, noise and marginal effects, CES preferences, expected income

	A. Aggregate estimation			B. Random coefficient model		
	a	r	μ	a	r	μ
Constant	0.798*** (0.112)	0.081 (0.393)	1.441*** (0.371)	0.618*** (0.085)	-0.105 (0.244)	0.587*** (0.116)
Expected income	0.040*** (0.021)	0.040 (0.115)	0.201*** (0.179)	0.015*** (0.006)	0.024 (0.017)	-0.006 (0.025)
Pre-clinical	0.073*** (0.053)	0.076* (0.168)	-0.123** (0.185)	0.109*** (0.033)	-0.006 (0.017)	0.006 (0.029)
Clinical	0.094*** (0.057)	0.149*** (0.195)	-0.052 (0.212)	0.155*** (0.039)	0.028 (0.019)	-0.077** (0.037)
Practical year	0.082*** (0.055)	0.183*** (0.245)	0.151 (0.372)	0.146*** (0.045)	0.056* (0.030)	-0.051 (0.034)
Female	-0.035*** (0.038)	-0.347*** (0.256)	0.070 (0.202)	-0.021 (0.019)	-0.058*** (0.021)	0.070* (0.038)
General altruism	-0.041*** (0.025)	-0.049*** (0.066)	-0.005 (0.050)	-0.044*** (0.010)	-0.091 (0.055)	0.173 (0.123)
N	693	693	693	693	693	693
Log-likelihood	-11,276.86	-11,276.86	-11,276.86	-5,216.73	-5,216.73	-5,216.73

Notes. This table shows the estimation results of the aggregate model and the random coefficient model for the CES preference functional with an expected income variable (=1 if expected income is above the median, =0 otherwise), study progress, gender, and general altruism in the set of covariates. The model also includes risk, time and social preferences and personality traits as covariates to account for observed heterogeneity. For the estimation results with the full list of covariates, see Tables D.6.1 - D.6.3 in Appendix D.6. Standard errors are clustered at the individual level. The lower number of observations compared to the full sample of $N = 733$ is due to subjects leaving the survey before completing the questionnaire. *p<0.10; **p<0.05; ***p<0.01.

specialties were regrouped as “Others” ($N = 305$, 41%).⁶⁶

Table 4.7 shows the estimation results of the aggregate estimation (Panel A) and of the random coefficient model (Panel B) when adding the speciality categories to the list of covariates. First, the effects of study progress, gender, and general altruism remain stable when controlling for specialty choices.⁶⁷ Second, stated preferences for specific specialties are linked to our altruism parameters. In our aggregate estimation, stating a preference for pediatrics or surgery relates to a significantly lower own-profit orientation a ; see left column of Panel A. For specializing in neurology/ psychiatry, a negative relation is only of weak statistical significance. When we additionally incorporate unobserved heterogeneity applying the random coefficient model, we find that the likelihood to state a preference for pediatrics and for surgery relates to a lower profit orientation; see left column of Panel B. Third, the aggregate estimation results suggest that inequity aversion tends to play a role for specialty selection. In particular, a preference for internal medicine, pediatrics or neurology/ psychiatry is linked to a decrease in parameter r ; see middle column of Panel A. This implies that medical students opting for these specialties express a lower tendency for inequity aversion. We, however, find no such link when estimating the random coefficient model.

4.6 Discussion and conclusion

The primary contribution of this paper is to the literature on eliciting preferences of (prospective) physicians. We also add to the more general economic literature on heterogeneity of social preferences using structural estimation techniques. In addition, some of our insights are relevant for several areas in health economics and also to the medical ethics and education literature that examines the role and development of altruism among healthcare professionals.

This study introduces a methodology to structurally estimate *patient-regarding altruistic preferences* of future physicians. In a novel experimental task, we measure the trade-off

⁶⁶For the full list of specialties, see Table D.6.4 in Appendix D.6.

⁶⁷Note that the same holds for other social and economic preferences and personality traits which are included in the list of covariates in the estimations but not explicitly reported in Table 4.7. The respective estimates are shown in Table D.6.5 and Table D.6.6 in Appendix D.6.

Table 4.7: Aggregate estimations and random coefficient model, preference parameters, noise and marginal effects, CES preferences

	A. Aggregate estimation			B. Random coefficient model		
	a	r	μ	a	r	μ
Constant	0.757*** (0.124)	-0.462*** (0.613)	2.157*** (0.624)	0.612*** (0.055)	-0.658* (0.388)	0.664*** (0.088)
Surgery	-0.027** (0.050)	0.069 (0.215)	-0.254*** (0.243)	-0.009* (0.005)	-0.014 (0.019)	0.017 (0.025)
Internal medicine	-0.010 (0.031)	0.201*** (0.187)	-0.536*** (0.234)	-0.017 (0.011)	-0.028 (0.029)	0.004 (0.029)
Pediatrics	-0.032*** (0.041)	0.298*** (0.267)	-0.464*** (0.271)	-0.015*** (0.005)	0.007 (0.015)	-0.035 (0.026)
Neurology/psychiatry	-0.020* (0.052)	0.317*** (0.272)	-0.182 (0.312)	-0.006 (0.008)	0.015 (0.015)	-0.066* (0.035)
Pre-clinical	0.083*** (0.048)	0.145** (0.263)	-0.279*** (0.290)	0.110*** (0.020)	0.002 (0.015)	-0.0001 (0.026)
Clinical	0.113*** (0.051)	0.307*** (0.256)	-0.159 (0.308)	0.168*** (0.021)	0.038** (0.017)	-0.086*** (0.032)
Practical year	0.062*** (0.071)	-0.084 (0.725)	0.834*** (0.910)	0.135*** (0.024)	0.054*** (0.019)	-0.024 (0.033)
Female	-0.049*** (0.041)	-0.489*** (0.335)	0.013 (0.260)	-0.050** (0.022)	-0.074*** (0.026)	0.069** (0.030)
General altruism	-0.053*** (0.021)	-0.129*** (0.098)	0.028 (0.064)	-0.043*** (0.010)	-0.053 (0.042)	0.152 (0.115)
N	729	729	729	729	729	729
Log-likelihood	-11,931.48	-11,931.48	-11,931.48	-5,522.58	-5,522.58	-5,522.58

Notes. This table shows the estimation results of the aggregate model and the random coefficient model for the CES preference functional with occupational choices, study progress, gender, and general altruism in the set of covariates. The model also includes other social and economic preferences and personality traits as covariates to account for observed heterogeneity. The lower number of observations compared to the full sample of $N = 733$ is due to subjects leaving the survey before completing the questionnaire. For the results of the full list of covariates, see Tables D.6.5 - D.6.7 in Appendix D.6. Standard errors are clustered at the individual level. * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$

between the physician's own profit and the patient benefit in a stylized physician-patient relationship for a large sample of medical students ($N = 733$). We find medical students to be rather altruistic by putting a weight of about two thirds on the patient health benefit and only one third on their own profit. In contrast, Li (2018), who studies altruistic preferences in a neutral context, finds that medical students put, on average, a weight of about two thirds on their own payoff and only one third on the payoff of an anonymous person. This difference highlights the importance of a distinction of physician altruism towards their patients and their altruism towards the general population. In comparison to our control group of non-medical students, our medical student sample behaves more patient-regarding overall, suggesting that altruistic motives already play an important role in the decision to enter medical school.

Our results further indicate that medical education affects patient-regarding altruistic preferences. Altruism is highest when starting medical education and significantly declines with study progress. Only at the end in the practical year patient-regarding altruism slightly increases again. One potential reason why patient-regarding altruism decreases could be a form of disillusionment of medical students. Students might realize a discrepancy between their own expectations and the professional reality when increasingly working in clinics and institutes during their medical education. The increase in patient-regarding altruism for practical-year students somewhat contradicts this reasoning. However, these students, who are about to graduate, represent a particular selective sample as they have successfully mastered the highest hurdles within the frame of medical education in Germany, namely the first and second part of the state exam. The selection process might have led to higher altruism in practical-year students. Moreover, medical education up to the practical year follows a rather strict, predefined curriculum. In the more theoretical oriented phases of medical education, patients might be rather considered as "learning objects". Whereas, the educational focus in the practical year lays on the interaction with patients. The more intense patient contact might lead to a reactivation of their own ideals prevalent prior to medical education. The National Competence-Based Learning Objectives Catalog of Medicine formulates training objectives in medical studies, based on targeted profiles of graduates with orientation to the physician role. It requires the deanery of medical schools

and policy-makers alike to set the course such that these objectives are worked towards or learned throughout medical education. As altruism plays a key aspect in the (benevolent) physician role, its teaching should be included in medical students' curriculum as advocated in the UK (Wicks et al., 2011). Also in Germany, the Medical German Education recommends that medical attitudes, ethical behavior and social skills should be taught during the course of study (Bundesärztekammer, 2020).

Our analyses also reveal that patient-regarding altruism are linked to subjects' characteristics, social preferences, income expectations, and specialty choices. Female medical students and students scoring higher in general altruism put significantly more weight on the patient's health benefit compared to their own profit. Less profit-oriented medical students also expect to earn less in their future when practicing as a physician. Finally, subjects who put a higher weight on the patient's health benefit are more likely to choose pediatrics and surgery as their preferred specialties.

One might argue that our results are sensitive to our choice of the structural model of altruistic preferences, the CES preference utility function. We found, however, that applying a Fehr and Schmidt (1999) preference functional, overall supports our main findings that patient-regarding altruistic preferences decrease with study progress. Medical students exhibit a high aversion to advantageous inequality. In our physician-patient context, this means that they as physicians are averse to gain a profit which is higher than the patients' benefits from the respective treatment choice. are reluctant to gaining more profit than the patient will benefit from their treatment choice. At the same time, students are somewhat (albeit less) averse to disadvantageous inequality. While clinical-phase students are the least concerned about advantageous inequality aversion, they exhibit the highest aversion to disadvantageous inequality.

We also add to a better understanding of how future physicians' altruistic preferences relate to their occupational choices. So far, experimental evidence has been inconclusive as to which extent expected earnings are behaviorally relevant for selecting into a particular specialty and whether other economic or non-economic factors might be more influential. We have formulated our concerns towards the approach Li (2018) used in her analysis of linking experimentally measured altruistic preferences to specialty choices of future physi-

cians. She categorizes US medical students' stated specialties into a low- or a high-income group by using a statistically defined cutoff point. One of our arguments was that it may be inappropriate to boil the analysis down to a decision between a group of high- and of low-income specialties. On the one hand, a physician's income is not solely determined by self-selection into a particular specialty. In Germany but also in other countries with publicly financed health care systems, a physician's income largely depends on many other factors, like the decision to work in a hospital or as resident physician, on regional differences, and the practice style which can - especially in the outpatient care sector - influence the earnings. On the other hand, various other non-monetary factors might influence the decision for a certain specialty, such as expected working hours (McKay, 1990), leisure time (Thornton and Esposto, 2003) and non-pecuniary attributes of work contracts (Holte et al., 2015). Given these findings, we included medical students' individual income expectations and specialty choices in the list of covariates. Our finding that less profit-oriented medical students also expect to earn less in the future when practicing as a physician admittedly could be sensitive to the way we constructed the expected income variable. Further analyses will be necessary to check for the robustness of our results.

There is an ongoing political debate on appropriate policy instruments to counterbalance current physician shortages. Statistics on physician supply in Germany, for instance, reveal high variations in the number of prospective physicians specializing in different medical fields as well as over time (Bundesärztekammer, 2019). Model calculations based on data from a nation-wide survey with medical students in Germany indicate that reoccupation rates can vary massively depending on the specialty (Jacob et al., 2019). For example, the modeled reoccupation rate for surgery as well as general medicine amounts to 54%, which means that only 54% of the current positions can be filled in future, and thus that every second positions would remain vacant. In contrast, the modeled rate for pediatrics amounts to 177%, which in turn means that about 77% more practical-year students want to become pediatricians than there are positions to fill. These developments can lead to insufficient supply in various medical specialties. In order to address the challenges effectively through policy interventions, a better understanding of the underlying motives of prospective physicians for selecting a specific specialty is essential.

While altruism is key in describing physician behavior, it is unclear whether self-selection into medical specialties is driven by patient-regarding altruism as well. Answers to our post-experimental questionnaire shed further light on a potential linkage, as we were able to show a stable positive relation between patient-regarding altruistic preferences of our medical student sample and their stated specialty choices. In particular, subjects who weigh the patient's health benefit higher, are more likely to select pediatrics and surgery as their preferred specialties. Since pediatrics includes the treatment of children and adolescents, who represent a particular vulnerable patient population, it is not unreasonable to assume that students with higher patient-regarding altruistic motives choose this specialty. Somewhat surprisingly, surgery is also linked to higher patient-regarding altruism. In surgery, however, the healing of patients is more visible and salient by the operations carried out than in other specialties which might relate to the higher weight on patients' health. Insofar, there seems to be a fit between our participants' patient-regarding motivations and selecting a specialty where such a motivation is particularly needed in the physician-patient relationship. Further, we also found a tendency that preferred specialty choices change in the course of medical education. The preference for pediatrics is comparatively stable from the beginning. Neurology/ psychiatry is decreasingly preferred and surgery substantially so, while the preference for internal and general medicine is increasing during the course of study; see also Table D.6.4 in Appendix D.6. The link between individuals' altruism and occupational choices is important from a health policy perspective, as medical education shapes preferences which then in turn have some predictive power to explain future physicians' labor market decisions. These important insights need to be studied in more detail in the future.

Bibliography

- Alexander, D. (2020). How do doctors respond to incentives? Unintended consequences of paying doctors to reduce costs. *Journal of Political Economy*, 128:4046–4096. doi:10.1086/710334.
- Allard, M., Jelovac, I., and Léger, P. T. (2011). Treatment and referral decisions under different physician payment mechanisms. *Journal of Health Economics*, 30:880–893. doi:10.1016/j.jhealeco.2011.05.016.
- Almlund, M., Duckworth, A. L., Heckman, J., and Kautz, T. (2011). Chapter 1 - Personality psychology and economics. in Hanushek, E.A., Machin, S., and Woessmann, L. (Eds.): *Handbook of the Economics of Education*, 4:1-181. Elsevier. doi.org/10.1016/B978-0-444-53444-6.00001-8.
- American Psychological Association (n.d.). *APA Dictionary of psychology*. Retrieved May 1, 2021, from <https://dictionary.apa.org/personality-trait>.
- Andreoni, J. (1989). Giving with impure altruism: Applications to charity and ricardian equivalence. *Journal of Political Economy*, 97:1447–1458. doi:10.1086/261662.
- Andreoni, J. and Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70:737–753. doi:10.1111/1468-0262.00302.
- Angerer, S., Glätzle-Rützle, D., and Waibel, C. (2021). Monitoring institutions in health care markets: Experimental evidence. *Health Economics*, 30:951–971. doi:10.1002/hec.4232.

- Anselmi, L., Borghi, J., Brown, G. W., Fichera, E., Hanson, K., Kadungure, A., Kovacs, R., Kristensen, S. R., Singh, N. S., and Sutton, M. (2020). Pay for performance: A reflection on how a global perspective could enhance policy and research. *International Journal of Health Policy and Management*, 9:365–369. doi:10.34172/ijhpm.2020.23.
- Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review*, 53:941–969. doi:10.1515/9780822385028-004.
- Ashraf, Nava, O. B. and Lee, S. S. (2016). Do-gooders and go-getters: Selection and performance in public service delivery. *International Growth Centre Working Paper*.
- Ashton, M. C. and Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, 91:340–345. doi:10.1080/00223890902935878.
- Baicker, K. and Goldman, D. (2011). Patient cost-sharing and healthcare spending growth. *Journal of Economic Perspectives*, 25:47–68. doi:10.1257/jep.25.2.47.
- Balafoutas, L. and Kerschbamer, R. (2020). Credence goods in the literature: What the past fifteen years have taught us about fraud, incentives, and the role of institutions. *Journal of Behavioral and Experimental Finance*, 26:100–285. doi:10.1016/j.jbef.2020.100285.
- Balgiu, B. (2018). The psychometric properties of the Big Five inventory-10 (bfi-10) including correlations with subjective and psychological well-being. *Global Journal of Psychology Research: New Trends and Issues*, 8:61–69.
- Barham, V. and Milliken, O. (2015). Payment mechanisms and the composition of physician practices: Balancing cost-containment, access, and quality of care. *Health Economics*, 24:895–906. doi:10.1002/hec.3069.
- Barros, P. and Braun, G. (2017). Upcoding in a national health service: The evidence from Portugal. *Health Economics*, 26:600–618. doi:10.1002/hec.3335.
- Bastani, H., Goh, J., and Bayati, M. (2019). Evidence of upcoding in pay-for-performance programs. *Management Science*, 65:1042–1060. doi:10.1287/mnsc.2017.2996.

- Bazzoli, G. J. (1985). Medical education indebtedness: Does it affect physician specialty choice? *Health Affairs*, 4:98–104. doi:10.1377/hlthaff.4.2.98.
- Beauchamp, T. L. and Childress, J. F. (2001). *Principles of biomedical ethics* (5th ed.). Oxford University Press.
- Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, 76:169–217.
- Ben-Ner, A., Putterman, L., Kong, F., and Magan, D. (2004). Reciprocity in a two-part dictator game. *Journal of Economic Behavior & Organization*, 53(3):333–352.
- Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96:1652–1678. doi:10.1257/aer.96.5.1652.
- Bexelius, T. S., Olsson, C., Järnbert-Pettersson, H., Parmskog, M., Ponzer, S., and Dahlin, M. (2016). Association between personality traits and future choice of specialisation among swedish doctors: a cross-sectional study. *Postgraduate Medical Journal*, 92:441–446. doi: 10.1136/postgradmedj-2015-133478.
- Borges, N. and Savickas, M. (2002). Personality and medical specialty choice: A literature review and integration. *Journal of Career Assessment*, 10:362–380. doi:10.1177/10672702010003006.
- Borghans, L., Duckworth, A. L., Heckman, J. J., and Ter Weel, B. (2008). The economics and psychology of personality traits. *Journal of Human Resources*, 43:972–1059. doi:10.3368/jhr.43.4.972.
- Bowles, S., Gintis, H., and Osborne, M. (2001). Incentive-enhancing preferences: Personality, behavior, and earnings. *American Economic Review*, 91(2):155–158.
- Brosig-Koch, J., Hennig-Schmidt, H., Kairies-Schwarz, N., Kokot, J., and Wiesen, D. (2020). Physician performance pay: Experimental evidence. HERO Online Working Paper Series 2020:3, University of Oslo, Health Economics Research Programme.

- Brosig-Koch, J., Hennig-Schmidt, H., Kairies-Schwarz, N., and Wiesen, D. (2016). Using artefactual field and lab experiments to investigate how fee-for-service and capitation affect medical service provision. *Journal of Economic Behavior & Organization*, 131, Part B:17–23. doi:10.1016/j.jebo.2015.04.011.
- Brosig-Koch, J., Hennig-Schmidt, H., Kairies-Schwarz, N., and Wiesen, D. (2017a). The effects of introducing mixed payment systems for physicians: Experimental evidence. *Health Economics*, 26:243–262. doi:10.1002/hec.3292.
- Brosig-Koch, J., Kairies-Schwarz, N., and Kokot, J. (2017b). Sorting into payment schemes and medical treatment: A laboratory experiment. *Health Economics*, 26:52–65. doi:10.1002/hec.3616.
- Bruhin, A., Fehr, E., and Schunk, D. (2019). The many faces of human sociality: Uncovering the distribution and stability of social preferences. *Journal of the European Economic Association*, 17:1025–1069. doi:10.1093/jea/jvy018.
- Bundesärztekammer (2019). *Ärztestatistik zum 31. Dezember 2019: Bundesgebiet gesamt*. https://www.bundesaerztekammer.de/fileadmin/user_upload/downloads/pdf-Ordner/Statistik2019/WEBStatistik_2019_k.pdf.
- Bundesärztekammer (2020). *Synopse Approbationsordnung für Ärzte (ÄApprO) aktuelle Fassung – Arbeitsentwurf Stellungnahme der Bundesärztekammer*. https://www.bundesaerztekammer.de/fileadmin/user_upload/downloads/pdf-Ordner/Stellungnahmen/AEApprO_Arbeitsentwurf_SN-BAEK_Synopse_final_24012020.pdf.
- Byambadalai, U., Ma, A., and Wiesen, D. (2019). Changing preferences: An experiment and estimation of market-incentive effects on altruism. Working paper, Boston University.
- Callen, M., Gulzar, S., Hasanain, A., Khan, M. Y., and Rezaee, A. (2018). Personalities and public sector performance: Evidence from a health experiment in Pakistan. *NBER Working Paper Series (21180)*. National Bureau of Economic Research.
- Campbell, S. M., Reeves, D., Kontopantelis, E., Sibbald, B., and Roland, M. (2009). Effects

- of pay for performance on the quality of primary care in England. *New England Journal of Medicine*, 361:368–378. doi:10.1056/NEJMsa0807651.
- Carciofo, R., Yang, J., Song, N., Du, F., and Zhang, K. (2016). Psychometric evaluation of Chinese-language 44-item and 10-item Big Five personality inventories, including correlations with chronotype, mindfulness and mind wandering. *PLOS ONE*, 11:1–26. doi.org/10.1371/journal.pone.0149963.
- Carter, G. M., Newhouse, J. P., and Relles, D. A. (1990). How much change in the case mix index is DRG creep? *Journal of Health Economics*, 9:411–428. doi:10.1016/0167-6296(90)90003-1.
- Chalkley, M. and Malcomson, J. M. (1998). Contracting for health services when patient demand does not reflect quality. *Journal of Health Economics*, 17:1–19. doi:10.1016/S0167-6296(97)00019-2.
- Chami, N. and Sweetman, A. (2019). Payment models in primary health care: A driver of the quantity and quality of medical laboratory utilization. *Health Economics*, 28:1166–1178. doi:10.1002/hec.3927.
- Chandra, A. and Skinner, J. (2012). Technology growth and expenditure growth in health care. *Journal of Economic Literature*, 50:645–680. doi:10.1257/jel.50.3.645.
- Charness, G. and Fehr, E. (2015). From the lab to the real world. *Science*, 350:512–513. doi:10.1126/science.aad4343.
- Choi, S., Fisman, R., Gale, D., and Kariv, S. (2007). Consistency and heterogeneity of individual behavior under uncertainty. *American Economic Review*, 97:1921–1938. doi:10.1257/aer.97.5.1921.
- Choné, P. and Ma, C. (2011). Optimal health care contract under physician agency. *Annals of Economics and Statistics*, 101/102:229–256. doi:10.2307/41615481.
- Clark, J. R. and Huckman, R. S. (2012). Broadening focus: Spillovers, complementarities, and specialization in the hospital industry. *Management Science*, 58:708–722. doi:10.1287/mnsc.1110.1448.

- Clemens, J. and Gottlieb, J. D. (2014). Do physicians' financial incentives affect medical treatment and patient health? *American Economic Review*, 104:1320–1349. doi:10.1257/aer.104.4.1320.
- Cox, J. C., Sadiraj, V., Schnier, K. E., and Sweeney, J. F. (2016a). Higher quality and lower cost from improving hospital discharge decision making. *Journal of Economic Behavior & Organization*, 131, Part B:1–16. doi:10.1016/j.jebo.2015.03.017.
- Cox, J. C., Sadiraj, V., Schnier, K. E., and Sweeney, J. F. (2016b). Incentivizing cost-effective reductions in hospital readmission rates. *Journal of Economic Behavior & Organization*, 131, Part B:24–35. doi:10.1016/j.jebo.2015.03.014.
- Crea, G., Galizzi, M. M., Linnosmaa, I., and Miraldo, M. (2019). Physician altruism and moral hazard: (no) evidence from Finnish national prescriptions data. *Journal of Health Economics*, 65:153–169. doi:10.1016/j.jhealeco.2019.03.006.
- Crede, M., Harms, P., Niehorster, S., and Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the Big Five personality traits. *Journal of Personality and social Psychology*, 102:874–888. doi:10.1037/a0027403.
- Cubel, M., Nuevo-Chiquero, A., Sánchez-Pagés, S., and Vidal-Fernandez, M. (2016). Do personality traits affect productivity? Evidence from the laboratory. *The Economic Journal*, 126:654–681. doi: 10.1111/ecoj.12373.
- Dafny, L. S. (2005). How do hospitals respond to price changes? *American Economic Review*, 95:1525–1547. doi:10.1257/000282805775014236.
- DellaVigna, S., List, J. A., and Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *Quarterly Journal of Economics*, 127:1–56. doi:10.1093/qje/qjr050.
- Di Guida, S., Gyrd-Hansen, D., and Oxholm, A. S. (2019). Testing the myth of fee-for-service and overprovision in health care. *Health Economics*, 28:717–722. doi:10.1002/hec.3875.

- Dohmen, T. and Falk, A. (2011). Performance pay and multidimensional sorting: Productivity, preferences, and gender. *American Economic Review*, 101:556–590. doi:10.1257/aer.101.2.556.
- Donato, K., Miller, G., Mohanan, M., Truskinovsky, Y., and Vera-Hernández, M. (2017). Personality traits and performance contracts: Evidence from a field experiment among maternity care providers in India. *American Economic Review*, 107:506–510. doi:10.1257/aer.p20171105.
- Doran, T., Fullwood, C., Gravelle, H., Reeves, D., Kontopantelis, E., Hiroeh, U., and Roland, M. (2006). Pay-for-performance programs in family practices in the United Kingdom. *New England Journal of Medicine*, 355:375–384. doi:10.1056/NEJMsa055505.
- Dorsey, E. R., David, J., and W, R. G. (2003). Influence of controllable lifestyle on recent trends in specialty choice by US medical students. *JAMA*, 290:1173–1178. doi:10.1001/jama.290.9.1173.
- Douven, R., Remmerswaal, M., and Zoutenbier, R. (2019). Do altruistic mental health care providers have better treatment outcomes? *Journal of Human Resources*, 54:310–341. doi:10.3368/jhr.54.2.0716.8070R1.
- Dufwenberg, M. and Dufwenberg, M. A. (2018). Lies in disguise – A theoretical analysis of cheating. *Journal of Economic Theory*, 175:248–264. doi:10.1016/j.jet.2018.01.013.
- Dulleck, U. and Kerschbamer, R. (2006). On doctors, mechanics, and computer specialists: The economics of credence goods. *Journal of Economic Literature*, 44:5–42. doi:10.1257/002205106776162717.
- Dulleck, U., Kerschbamer, R., and Sutter, M. (2011). The economics of Credence Goods: An experiment on the role of liability, verifiability, reputation, and competition. *American Economic Review*, 101:526–555. doi:10.1257/aer.101.2.526.
- Eijkenaar, F., M. Emmert, M. Scheppach, and Oliver Schoeffski (2013). Effects of pay for performance in health care: A systematic review of systematic reviews. *Health Policy*, 110:115–130. doi:10.1016/j.healthpol.2013.01.008.

- Eilermann, K., Halstenberg, K., Kuntz, L., Martakis, K., Roth, B., and Wiesen, D. (2019). The effect of expert feedback on antibiotic prescribing in pediatrics: Experimental evidence. *Medical Decision Making*, 39:781–795. doi:10.1177/0272989X19866699.
- Eisinga, R., Grotenhuis, M., and Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health*, 58:637–642. doi:10.1007/s00038-012-0416-3.
- Ellis, R. P. and McGuire, T. G. (1986). Provider behavior under prospective reimbursement: Cost sharing and supply. *Journal of Health Economics*, 5:129–151. doi:10.1016/0167-6296(86)90002-0.
- Ellis, R. P. and McGuire, T. G. (1990). Optimal payment systems for health services. *Journal of Health Economics*, 9:375–396. doi:10.1016/0167-6296(90)90001-J.
- Emmert, M., Eijkenaar, F., Kemter, H., Esslinger, A., and Schöffski, O. (2012). Economic evaluation of pay-for-performance in health care: A systematic review. *European Journal of Health Economics*, 13:755–767. doi:10.1007/s10198-011-0329-8.
- Epstein, A. M. (2012). Will pay for performance improve quality of care? The answer is in the details. *New England Journal of Medicine*, 367:1852–1853. doi:10.1056/NEJMe1212133.
- Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. (2018). Global evidence on economic preferences. *Quarterly Journal of Economics*, 133:1645–1692. doi:10.1093/qje/qjy013.
- Falk, A., Becker, A., Dohmen, T. J., Huffman, D., and Sunde, U. (2016). The preference survey module: A validated instrument for measuring risk, time, and social preferences. *IZA Discussion Paper No. 9674*.
- Falk, A. and Heckman, J. J. (2009). Lab experiments are a major source of knowledge in social sciences. *Science*, 326:535–538. doi:10.1126/science.1168244.
- Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114:817–868. doi:10.1162/003355399556151.

- Fischbacher, U. (2007). z-Tree: Zurich toolbox for readymade economic experiments – Experimenter’s manual. *Experimental Economics*, 10:171–178. doi:10.1007/s10683-006-9159-4.
- Fisman, R., Kariv, S., and Markovits, D. (2007). Individual preferences for giving. *American Economic Review*, 97:1858–1876. doi:10.1257/aer.97.5.1858.
- Flaherman, V. J., Schaefer, E. W., Kuzniewicz, M. W., Li, S. X., Walsh, E. M., and Paul, I. M. (2015). Early weight loss nomograms for exclusively breastfed newborns. *Pediatrics*, 135:16–23. doi:10.1542/peds.2014-1532.
- Fletcher, J. M. (2013). The effects of personality traits on adult labor market outcomes: Evidence from siblings. *Journal of Economic Behavior & Organization*, 89:122–135. doi:10.1016/j.jebo.2013.02.004.
- Fulmer, I. and Walker, W. (2015). More bang for the buck?: Personality traits as moderators of responsiveness to pay-for-performance. *Human Performance*, 28:40–65. doi:10.1080/08959285.2014.974755.
- Furnham, A. (2008). Relationship among four Big Five measures of different length. *Psychological Reports*, 102:312–316. doi:10.2466/pr0.102.1.312-316.
- Gagné, R. and Léger, P. T. (2005). Determinants of physicians’ decisions to specialize. *Health Economics*, 14:721–735. doi:10.1002/hec.970.
- Galizzi, M. M., Tammi, T., Godager, G., Linnosmaa, I., and Wiesen, D. (2015). Provider altruism in health economics. *THL Discussion paper 4/2015*, National Institute for Health and Welfare.
- Galizzi, M. M. and Wiesen, D. (2017). Behavioural experiments in health: An introduction. *Health Economics*, 26:3–5. doi:10.1002/hec.3629.
- Galizzi, M. M. and Wiesen, D. (2018). Behavioral experiments in health economics. In Hamilton, J., editor, *Oxford Research Encyclopedia of Economics and Finance*. Oxford University Press, Oxford, UK. doi:10.1093/acrefore/9780190625979.013.244.

- Glazier, R. H., Klein-Geltink, J., Kopp, A., and Sibley, L. M. (2009). Capitation and enhanced fee-for-service models for primary care reform: A population-based evaluation. *CMAJ: Canadian Medical Association journal*, 180:E72–E81. doi:10.1503/cmaj.081316.
- Gneezy, U., Kajackaite, A., and Sobel, J. (2018). Lying aversion and the size of the lie. *American Economic Review*, 108:419–53. doi:10.1257/aer.20161553.
- Godager, G., Hennig-Schmidt, H., and Iversen, T. (2016). Does performance disclosure influence physicians’ medical decisions? An experimental study. *Journal of Economic Behavior & Organization*, 131:36–46. doi:10.1016/j.jebo.2015.10.005.
- Godager, G. and Wiesen, D. (2013). Profit or patients’ health benefit? Exploring the heterogeneity in physician altruism. *Journal of Health Economics*, 32:1105–1116. doi:10.1016/j.jhealeco.2013.08.008.
- Gosling, S. D., Rentfrow, P. J., and Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37:504–528. doi:10.1016/S0092-6566(03)00046-1.
- Gravelle, H., Sutton, M., and Ma, A. (2010). Doctor behaviour under a pay for performance contract: Treating, cheating and case finding? *Economic Journal*, 120:129–156. doi:10.1111/j.1468-0297.2009.02340.x.
- Green, E. P. (2014). Payment systems in the healthcare industry: An experimental study of physician incentives. *Journal of Economic Behavior & Organization*, 106:367–378. doi:10.1016/j.jebo.2014.05.009.
- Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1:114–125. doi:10.1007/s40881-015-0004-4.
- Greiner, B., Zhang, L., and Tang, C. (2017). Separation of prescription and treatment in health care markets: A laboratory experiment. *Health Economics*, 26:21–35. doi:10.1002/hec.3575.

- Groß, M., Jürges, H., and Wiesen, D. (2021). The effects of audits and fines on upcoding in neonatology. *Health Economics*, 30:1978–1986. doi:10.1002/hec.4272.
- Harris, M. G., Gavel, P. H., and Young, J. R. (2005). Factors influencing the choice of specialty of Australian medical graduates. *Medical Journal of Australia*, 31:813–823. doi:10.5694/j.1326–5377.2005.tb07058.x.
- Hellerstein, J. K. (1998). The importance of the physician in the generic versus trade-name prescription decision. *Rand Journal of Economics*, 29:108–136. doi: 10.2307/2555818.
- Hennig-Schmidt, H., Jürges, H., and Wiesen, D. (2019). Dishonesty in health care practice: A behavioral experiment on upcoding in neonatology. *Health Economics*, 28:319–338. doi:10.1002/hec.3842.
- Hennig-Schmidt, H., Selten, R., and Wiesen, D. (2011). How payment systems affect physicians’ provision behavior – An experimental investigation. *Journal of Health Economics*, 30:637–646. doi:10.1016/j.jhealeco.2011.05.001.
- Hennig-Schmidt, H. and Wiesen, D. (2014). Other-regarding behavior and motivation in health care provision: An experiment with medical and non-medical students. *Social Science & Medicine*, 108:156 – 165. doi:10.1016/j.socscimed.2014.03.001.
- Herbst, D. and Mas, A. (2015). Peer effects on worker output in the laboratory generalize to the field. *Science*, 350:545–549. doi:10.1126/science.aac9555.
- Hochuli, P. (2020). Losing body weight for money: How provider-side financial incentives cause weight loss in Swiss low-birth-weight newborns. *Health Economics*, 29:406–418. doi:10.1002/hec.3991.
- Hole, A. R. (2015). DCREATE: Stata module to create efficient designs for discrete choice experiments. *Statistical Software Components*, Boston College Department of Economics.
- Holte, J. H., Kjaer, T., Abelsen, B., and Olsen, J. A. (2015). The impact of pecuniary and non-pecuniary incentives for attracting young doctors to rural general practice. *Social Science & Medicine*, 128:1–9. doi:10.1016/j.socscimed.2014.12.022.

- Huck, S., Lünser, G., Spitzer, F., and Tyran, J.-R. (2016). Medical insurance and free choice of physician shape patient overtreatment: A laboratory experiment. *Journal of Economic Behavior & Organization*, 131:78–105. doi:10.1016/j.jebo.2016.06.009.
- Huesmann, K., Waibel, C., and Wiesen, D. (2020). Rankings in health care organizations. *Working Paper*. doi:10.2139/ssrn.3690851.
- Iversen, T. and Lurås, H. (2000). The effect of capitation on GPs’ referral decisions. *Health Economics*, 9:199–210. doi:10.1002/(sici)1099-1050(200004)9:3<199::aid-hec514>3.0.co;2-2.
- Jack, W. (2005). Purchasing health care services from providers with unknown altruism. *Journal of Health Economics*, 24:73–93. doi:10.1016/j.jhealeco.2004.06.001.
- Jacob, R., Kopp, J., and Felling, P. (2019). *Berufsmonitoring Medizinstudenten 2018: Ergebnisse einer bundesweiten Befragung. Kassenärztliche Bundesvereinigung*. https://www.kbv.de/media/sp/Berufsmonitoring_Medizinstudierende_2018.pdf.
- Januleviciute, J., Askildsen, J. E., Kaarboe, O., Siciliani, L., and Sutton, M. (2016). How do hospitals respond to price changes? Evidence from Norway. *Health Economics*, 25:620–636. doi:10.1002/hec.3179.
- Jauhar, S. (2014). *Doctored: The Disillusionment of an American Physician*. Farrar, Straus and Giroux, New York.
- Jia, L., Meng, Q., Scott, A., Yuan, B., and Zhang, L. (2021). Payment methods for healthcare providers working in outpatient healthcare settings. *Cochrane Database of Systematic Reviews*, 1:CD011865. doi:10.1002/14651858.CD011865.pub2.
- John, O., Naumann, L., and Soto, C. (2008). Paradigm shift to the integrative big five taxonomy. In John, O.P., Robins, R.W., and Pervin, L.A. (Eds.): *Handbook of personality: Theory and research* (3rd ed.), 3:114–158. Guilford Press.
- Jürges, H. and Köberlein, J. (2015). What explains DRG upcoding in neonatology? The roles of financial incentives and infant health. *Journal of Health Economics*, 43:13–26. doi:10.1016/j.jhealeco.2015.06.001.

- Kerschbamer, R. and Sutter, M. (2017). The economics of credence goods – A survey of recent lab and field experiments. *CESifo Economic Studies*, 63:1–23. doi:10.1093/cesifo/ifx001.
- Keser, C., Peterle, E., and Schnitzler, C. (2014). Money talks-Paying physicians for performance. *Cege Discussion Paper*, University of Göttingen, 173. doi:10.2139/ssrn.2357326.
- Kesternich, I., Schumacher, H., and Winter, J. (2015). Professional norms and physician behavior: Homo oeconomicus or homo hippocraticus? *Journal of Public Economics*, 131:1–11. doi:10.1016/j.jpubeco.2015.08.009.
- Kolstad, J. T. (2013). Information and quality when motivation is intrinsic: Evidence from surgeon report cards. *American Economic Review*, 103:2875–2910. doi:10.1257/aer.103.7.2875.
- Kristensen, S. R., Siciliani, L., and Sutton, M. (2016). Optimal price-setting in pay for performance schemes in health care. *Journal of Economic Behavior & Organization*, 123:57–77. doi:10.1016/j.jebo.2015.12.002.
- Lagarde, M. and Blaauw, D. (2017). Physicians’ responses to financial and social incentives: A medically framed real effort experiment. *Social Science & Medicine*, 179:147–159. doi:10.1016/j.socscimed.2017.03.002.
- Li, J. (2018). Plastic surgery or primary care? Altruistic preferences and expected specialty choice of U.S. medical students. *Journal of Health Economics*, 62:45–59. doi:10.1016/j.jhealeco.2018.09.005.
- Li, J., Dow, W. H., and Kariv, S. (2017). Social preferences of future physicians. *Proceedings of the National Academy of Sciences*, 114:10291–10300. doi:10.1073/pnas.1705451114.
- Li, J., Hurley, J., DeCicca, P., and Buckley, G. (2014). Physician response to pay-for-performance: Evidence from a natural experiment. *Health Economics*, 23:962–978. doi:10.1002/hec.2971.
- Lindenauer, P. K., Remus, D., Roman, S., Rothberg, M. B., Benjamin, E. M., Ma, A., and Bratzler, D. W. (2007). Public reporting and pay for performance in hospital quality improvement. *New England Journal of Medicine*, 356:486–496. doi:10.1056/NEJMsa064964.

- Liu, T. and Ma, C. A. (2013). Health insurance, treatment plan, and delegation to altruistic physician. *Journal of Economic Behavior & Organization*, 85:79–96. doi:10.1016/j.jebo.2012.11.002.
- Lundin, D. (2000). Moral hazard in physician prescription behavior. *Journal of Health Economics*, 19:639–662. doi:10.1016/S0167-6296(00)00033-3.
- Ma, C. A. (1994). Health care payment systems: Cost and quality incentives. *Journal of Economics and Management Strategy*, 3:93–112. doi:10.1111/j.1430-9134.1994.00093.x.
- Marburger Bund (n.d.). *Tarifverträge*. Retrieved May 19, 2021, from <https://www.marburger-bund.de/bundesverband/tarifvertraege>.
- Martinsson, P. and Persson, E. (2019). Physician behavior and conditional altruism: The effects of payment system and uncertain health benefit. *Theory and Decision*, 87:365–387. doi:10.1007/s11238-019-09714-7.
- Mathes, T., Pieper, D., Morche, J., Polus, S., Jaschinski, T., and M., E. (2019). Pay for performance for hospitals. *Cochrane Database of Systematic Reviews*, 7. doi:10.1002/14651858.CD011156.pub2.
- Maynard, A. (2012). The powers and pitfalls of payment for performance. *Health Economics*, 21:3–12. doi:10.1002/hec.1810.
- McFadden, D. (1981). Econometric models of probabilistic choice. In Manski, C. and McFadden, D. (Eds.): *Structural analysis of discrete data with econometric applications*, 198-272. MIT Press.
- McGuire, T. G. (2000). Physician agency. In Cuyler, A. J. and Newhouse, J. P. (Eds.): *Handbook of Health Economics*, 1, Part A:461-536. Elsevier. doi:10.1016/s1574-0064(00)80168-7.
- McKay, N. L. (1990). The economic determinants of specialty choice by medical residents. *Journal of Health Economics*, 9:335–357. doi:10.1016/0167-6296(90)90050-d.
- Meacock, R., Kristensen, S. R., and Sutton, M. (2014). The cost-effectiveness of using

- financial incentives to improve provider quality: A framework and application. *Health Economics*, 23:1–13. doi:10.1002/hec.2978.
- Mendelson, A., Kondo, K., Damberg, C., Low, A., Motuapuaka, M., Freeman, M., O’Neil, M., Relevo, R., and Kansagara, D. (2017). The effects of pay-for-performance programs on health, health care use, and processes of care: A systematic review. *Annals of Internal Medicine*, 166:doi:10.7326/M16–1881.
- Milstein, R. and Schreyögg, J. (2016). Pay for performance in the inpatient sector: A review of 34 P4P programs in 14 OECD countries. *Health Policy*, 120:1125–1140. doi:10.1016/j.healthpol.2016.08.009.
- Mimra, W., Rasch, A., and Waibel, C. (2016). Second opinions in markets for expert services: Experimental evidence. *Journal of Economic Behavior & Organization*, 131, Part B:106–125. doi:10.1016/j.jebo.2016.03.004.
- Mullen, K., Frank, R., and Rosenthal, M. (2010). Can you get what you pay for? Pay-for-performance and the quality of healthcare providers. *RAND Journal of Economics*, 41:64–91. doi:10.1111/j.1756–2171.2009.00090.x.
- Müller, J. and Schwieren, C. (2012). Can personality explain what is underlying women’s unwillingness to compete? *Journal of Economic Psychology*, 33:448–460. doi:10.1016/j.joep.2011.12.005.
- Mullola, S., Hakulinen, C., Presseau, J., Gimeno Ruiz de Porras, D., Jokela, M., Hintsala, T., and Elovainio, M. (2018). Personality traits and career choices among physicians in finland: employment sector, clinical patient contact, specialty and change of specialty. *BMC Medical Education*, 18(1):52. doi:10.1186/s12909–018–1155–9.
- Nicholson, S. (2002). Physician specialty choice under uncertainty. *Journal of Labor Economics*, 20:816–847. doi:10.1086/342039.
- Nicholson, S. and Propper, C. (2011). Medical workforce. In Pauly, M. V., McGuire, T. G., and Barros, P. P. (Eds.): *Handbook of Health Economics*, 2:198–272. Elsevier. doi:10.1016/b978-0-444-53592-4.00014-1.

- Nyhus, E. K. and Pons, E. (2005). The effects of personality on earnings. *Journal of Economic Psychology*, 26:363–384. doi:10.1016/j.joep.2004.07.001.
- Olivella, P. and Siciliani, L. (2017). Reputational concerns with altruistic providers. *Journal of Health Economics*, 55:1–13. doi:10.1016/j.jhealeco.2017.05.003.
- Oxholm, A.-S., Di Guida, S., and Gyrd-Hansen, D. (2021). Allocation of health care under pay for performance: Winners and losers. *Social Science & Medicine*, 278:113939. doi:10.1016/j.socscimed.2021.113939.
- Peckham, S. and Wallace, A. (2010). Pay for performance schemes in primary care: What have we learnt? *Quality in Primary Care*, 18:111–116.
- Pellegrino, E. D. (1987). Altruism, self-interest, and medical ethics. *JAMA*, 258:1939–1940. doi:10.1001/jama.1987.03400140101036.
- Rammstedt, B. and John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41:203–212. doi:10.1016/j.jrp.2006.02.001.
- Reif, S., Hafner, L., and Seebauer, M. (2020). Physician behavior under prospective payment schemes—Evidence from artefactual field and lab experiments. *International Journal of Environmental Research and Public Health*, 17:5540. doi:10.3390/ijerph17155540.
- Reif, S., Wichert, S., and Wuppermann, A. (2018). Is it good to be too light? Birth weight thresholds in hospital reimbursement systems. *Journal of Health Economics*, 59:1–25. doi:10.1016/j.jhealeco.2018.01.007.
- Roland, M. (2004). Linking physicians’ pay to the quality of care - A major experiment in the United Kingdom. *New England Journal of Medicine*, 351:1448–1454. doi:10.1056/NEJMhpr041294.
- Roland, M. and Campbell, S. (2014). Successes and failures of pay for performance in the United Kingdom. *New England Journal of Medicine*, 370:1944–1949. doi:10.1056/NEJMhpr1316051.

- Roland, M. and Olesen, F. (2015). Can pay for performance improve the quality of primary care. *BMJ (Clinical research ed.)*, 354. doi:10.1136/bmj.i4058.
- Rosenthal, M. B., Landon, B. E., Normand, S.-L. T., Frank, R. G., and Epstein, A. M. (2006). Pay for performance in commercial HMOs. *New England Journal of Medicine*, 355:1895–1902. doi:10.1056/NEJMsa063682.
- Roth, A. E. (2002). The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica*, 70:1341–1378. doi:10.1111/1468-0262.00335.
- Schlenker, B. R. (2008). Integrity and character: Implications of principled and expedient ethical ideologies. *Journal of Social and Clinical Psychology*, 27:1078–1125. doi:10.1521/jscp.2008.27.10.1078.
- Scott, A., Sivey, P., Ouakrim, D. A., Willenberg, L., Naccarella, L., Furler, J., and Young, D. (2011). The effect of financial incentives on the quality of health care provided by primary care physicians. *Cochrane Database of Systematic Reviews*, CD008451. doi:10.1002/14651858.CD008451.pub2.
- Shigeoka, H. and Fushimi, K. (2014). Supplier-induced demand for newborn treatment: Evidence from Japan. *Journal of Health Economics*, 35:162–178. doi:10.1016/j.jhealeco.2014.03.003.
- Silverman, E. and Skinner, J. (2004). Medicare upcoding and hospital ownership. *Journal of Health Economics*, 23:369–389. doi:10.1016/j.jhealeco.2003.09.007.
- Sivey, P., Scott, A., Witt, J., Joyce, C., and Humphreys, J. (2012). Junior doctors’ preferences for specialty choice. *Journal of Health Economics*, 31:813–823. doi:10.1016/j.jhealeco.2012.07.001.
- Snowberg, E. and Yariv, L. (2021). Testing the waters: Behavior across participant pools. *American Economic Review*, 111:687–719. doi:10.1257/aer.20181065.

- Song, Z., Ji, Y., Safran, D. G., and Chernew, M. E. (2019). Health care spending, utilization, and quality 8 years into global payment. *New England Journal of Medicine*, 381:252–263. doi:10.1056/NEJMsa1813621.
- Statistisches Bundesamt (2018a). *Bildung und Kultur- Prüfungen an Hochschulen 2017*. Fachserie 11 Reihe 4.2. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Hochschulen/Publikationen/Downloads-Hochschulen/pruefungen-hochschulen-2110420177004.pdf?__blob=publicationFile&v=4.
- Statistisches Bundesamt (2018b). *Kostenstruktur bei Arzt- und Zahnarztpraxen sowie Praxen von psychologischen Psychotherapeuten 2015*. Fachserie 2 Reihe 1.6.1. https://www.destatis.de/DE/Themen/Branchen-Unternehmen/Dienstleistungen/Publikationen/Downloads-Dienstleistungen-Kostenstruktur/kostenstruktur-aerzte-2020161159004.pdf;jsessionid=D942E64F7A9DBE3A447CE4ECA242613E.live721?__blob=publicationFile.
- Statistisches Bundesamt (2021). *Studierende insgesamt und Studierende Deutsche im Studienfach Medizin (Allgemein-Medizin) nach Geschlecht*. Retrieved January 7, 2021, from <https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Hochschulen/Tabellen/lrbil05.html>.
- Stokes, J., Struckmann, V., Kristensen, S. R., Fuchs, S., van Ginneken, E., Tsiachristas, A., Rutten van Mölken, M., and Sutton, M. (2018). Towards incentivising integration: A typology of payments for integrated care. *Health Policy*, 122:963–969. doi:10.1016/j.healthpol.2018.07.003.
- Thalmayer, A., Saucier, G., and Eigenhuis, A. (2011). Comparative validity of brief to medium-length big five and big six personality questionnaires. *Psychological Assessment*, 23:995–1009. doi:10.1037/a0024165.
- Thornton, J. and Esposto, F. (2003). How important are economic factors in choice of medical specialty? *Health Economics*, 12:67–73. doi:10.1002/hec.682.

- Train, K. E. (2009). *Discrete choice methods with simulation* (2nd ed.). Cambridge University Press.
- Waibel, C. and Wiesen, D. (2021). An experiment on referrals in health care. *European Economic Review*, 131:103612. doi:10.1016/j.euroecorev.2020.103612.
- Wang, J., Iversen, T., Hennig-Schmidt, H., and Godager, G. (2020). Are patient-regarding preferences stable? Evidence from a laboratory experiment with physicians and medical students from different countries. *European Economic Review*, 125:103411. doi:10.1016/j.euroecorev.2020.103411.
- Wicks, L., Noor, S., and Rajaratnam, V. (2011). Altruism and medicine. *BMJ*, 343:d4537. doi:10.1136/bmj.d4537.
- World Medical Association (2018). *WMA Declaration of Geneva: The Physician's Pledge*. Retrieved June 25, 2021, from <https://www.wma.net/policies-post/wma-declaration-of-geneva/>.
- Zavlin, D., Jubbal, K., Noe, J., and Gansbacher, B. (2017). A comparison of medical education in Germany and the United States: From applying to medical school to the beginnings of residency. *German Medical Science*, 15. doi:10.3205/000256.
- Zentralinstitut für die Kassenärztliche Versorgung (2016). *Zi-Praxis-Panel: Jahresbericht 2016 - Wirtschaftliche Situation und Rahmenbedingungen in der vertragsärztlichen Versorgung der Jahre 2012 bis 2015*. 7. Jahrgang. https://www.zi-pp.de/pdf/ZiPP_Jahresbericht_2016.pdf.

Appendices

A Appendix to Chapter 1

A.1 Additional information about the experiment

A.1.1 Experiment instructions

Instructions for the baseline treatment (NANF), translated from German. [The text in brackets is for the treatments 10ANF, 10AF, and 75AF.]

Description of the experiment

General information

You are participating in an economic decision-making experiment. Please read the experiment description carefully. It is very important that **you do not talk with other participants for the entire duration of the experiment**. If you violate this rule, you will be excluded from the experiment and will not receive any payment.

If there is anything you do not understand, please take another look at this description of the experiment. If you still have questions, please raise your hand. We will come to your cubicle and answer your question personally.

You can earn money in the course of the experiment. The amount of your earnings depends on your decisions. All monetary amounts are displayed in ‘Taler’, at the following rate:

$$1 \text{ Taler} = 1 \text{ Euro Cent.}$$

At the end of the experiment, you will receive the payment you have earned in cash.

You will make your decisions on a computer screen in your cubicle. Your decisions cannot be attributed to you personally. All data and answers will therefore be analyzed anonymously.

Decision situation

The experiment is based on a decision situation in the obstetrics practice of a hospital's large neonatal care unit. During the experiment, you are entrusted with recording the weight of very early-born infants (birth weights). All participants in this experiment are confronted with the same decision situation.

You have the task of **recording different birth weights, which are displayed on a scale, into a birth report**. The recorded birth weight determines the lump-sum reimbursement per case, which comprises one part of your payment for the experiment.

In total, **six birth weights of different early-born infants will be displayed on the scale, in grams (g)**, on your computer screen. The birth weights displayed can have values between 1,150g and 1,550g. Each birth weight appears only once.

You can record the birth weights in the report in 50-gram intervals, i.e., **1,150, 1,200, 1,250, 1,300, 1,350, 1,400, 1,450, 1,500, or 1,550**.

The birth weight recorded in the report determines which lump-sum reimbursement (in Taler) will be paid for the treatment of an early-born infant. The medical treatment of the early-born infants incurs costs, which should, on average, be compensated by the the lump-sum reimbursement. The costs are dependent on the birth weight displayed on the scale. It is provided for that the early-born infants receive optimal medical care in the neonatal care unit, irrespectively of the lump-sum reimbursement paid.

Lump-sum reimbursements and costs by birth weight are as follows:

Birth weight (in g)	1,150	1,200	1,250	1,300	1,350	1,400	1,450	1,500	1,550
Lump-sum reimbursement (in Taler)	380	380	200	200	200	200	200	120	120
Costs (in Taler)	340	300	260	230	200	180	160	140	130

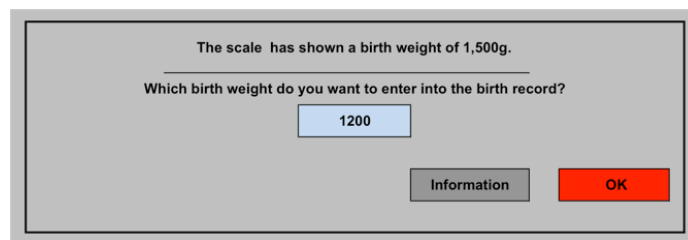
[The newborn will be weighed again the day after birth. This weight will also be recorded in the birth report, by a different person. It should be noted that a newborn baby loses up to 5% of its birth weight in the first 24 hours.

In order to audit hospital bills, the health insurer randomly compares the weights you recorded with the weights measured on the day after the baby's birth. This audit occurs **after** all birth weights have been recorded **with a probability of 10%** [10ANF and 10AF], **75%** [75AF].

If the weight measured on the day after birth is **higher than the recorded birth weight**, the health insurer assumes that the recorded birth weight was too low. This constitutes a serious fake entry.

Example: Given a birth weight of 1,500g, the weight 24 hours after birth is at least 1,425g. If the recorded birth weight is, for example, 1,400g, then the health insurer assumes a serious fake entry since the weight on the day after birth is in fact over 1,400g.]

You make your decisions about what birth weight to record individually and anonymously on your computer screen. An example of the decision screen is found in the following screenshot:



The screenshot shows a computer interface for recording birth weight. At the top, it says "The scale has shown a birth weight of 1,500g." Below this, a question asks "Which birth weight do you want to enter into the birth record?". There is a text input field containing the number "1200". At the bottom, there are two buttons: a grey "Information" button and a red "OK" button.

You enter the birth weight in the designated field. After you have entered the birth weight, you have the opportunity to learn about the costs arising from the medical care of the early-born infant [and about the minimal weight on the day after birth]. To access this, click on the 'Information' button. The following screenshot provides an example:

You confirm your decision by clicking on 'OK' and continue to the next screen.

The scale has shown a birth weight of 1,500g.

Which birth weight do you want to enter into the birth record?

1200

Information OK

For the birth weight you entered the following applies:

Weight:	1,200g
Lump-sum reimbursement:	380
Costs for medical treatment:	140
Payment:	240

When weighing the day after birth, the weight is at least:
1,425g

[The left panel shows the screen for the no-audit-no-fine treatment, the right panel for the audit treatments.]

Payments in the experiment

Your payment for each recorded birth weight is determined as follows:

Lump-sum reimbursement per case according to the recorded birth weight

minus

Costs for treatment of the early-born infant according to the birth weight displayed on the scale.

After you have made your six decisions, your payment for all entries is calculated as the sum of all six decisions.

Additionally, you will receive a **fixed amount of 400 Taler**. Your total payment is therefore comprised of your payment for the recorded birth weights and the fixed amount. [In the event of the insurance provider's audit showing that the weight on the day after birth **for at least one** newborn baby is larger than the recorded weight, this constitutes a serious fake entry. You will therefore be punished with a **reduction in your total payment equaling the amount of your payments for all recorded birth weights, and you will receive the fixed amount only**. This means that the hospital has to treat all children without lump-sum reimbursements. It is not known at the time the treatment is provided whether or not this is the case and therefore has no effect on medical care.

You will learn at the end of the experiment whether the birth weights you entered were audited and whether a fake entry was discovered.]

Your total payment for the experiment will be converted into Euro and paid out to you at the end of the experiment.

Before the actual experiment begins, we would like to ask you to answer several practice questions that should help you to understand the experiment better.

After you have made all of your decisions, we would ask you to answer a series of questions contained in a questionnaire.

A.1.2 Parameters in the experiment

Table A.1.1: Profit matrix

True birth weight w_j (in g)	Reported birth weight \hat{w}_j (in g)								
	1150	1200	1250	1300	1350	1400	1450	1500	1550
1200	80	80	-100	-100	-100	-100	-100	-180	-180
1250	120	120	-60	-60	-60	-60	-60	-140	-140
1300	150	150	-30	-30	-30	-30	-30	-110	-110
1350	180	180	0	0	0	0	0	-80	-80
1400	200	200	20	20	20	20	20	-60	-60
1500	240	240	60	60	60	60	60	-20	-20

Notes. This table shows profits for the weights in the subjects' choice range (first column) and the birth weights that can be entered in the birth report (second row). Note that reimbursements depend on the reported birth weights. DRG thresholds are at 1,250 and 1,500. Costs depend on the true birth weight and increase with decreasing birth weights. All monetary amounts are given in Taler, our experimental currency, the exchange rate being 1 Taler = 0.01 EUR.

A.1.3 Decomposition of the effects of audits and fines

In our experiment, we observe four of six possible combination of detection probabilities and fines:

Table A.1.2: Decomposition of the effects of audits and fines

Detection probability	No fine	Fine	Δ
0% (no audit)	\bar{y}_{00} (NANF)	—	
10%	\bar{y}_{10} (10ANF)	\bar{y}_{11} (10AF)	$\bar{y}_{11} - \bar{y}_{10}$
75%	—	\bar{y}_{21} (75AF)	
Δ	$\bar{y}_{10} - \bar{y}_{00}$	$\bar{y}_{21} - \bar{y}_{11}$	

Assuming that participants are randomly allocated to a cell, NANF can serve as the counterfactual out- come for 10ANF, and 10AF as the counterfactual outcome for 75AF. Hence, for the effect of audits, we consider two vertical comparisons:

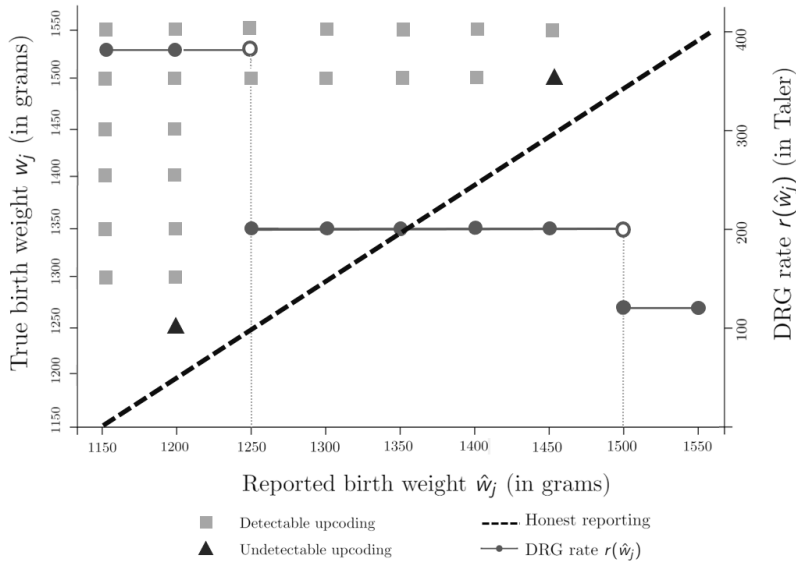
- $\Delta(10\text{ANF}, \text{NANF}) = \bar{y}_{10} - \bar{y}_{00}$: for the introduction of an audit with small detection probability (10%) without a fine
- $\Delta(75\text{AF}, 10\text{AF}) = \bar{y}_{21} - \bar{y}_{11}$: for a substantial increase in the detection probability, i.e., raising the detection probability from small (10%) to large (75%)

For the effect of fines, we consider the horizontal comparisons:

- $\Delta(10\text{AF}, 10\text{ANF}) = \bar{y}_{11} - \bar{y}_{10}$: for the introduction of a fine at a 10%-detection probability

A.1.4 Graphical explanation of detectable and undetectable upcoding

Figure A.1.1: Graphical explanation of upcoding



A.1.5 Further details on the experimental procedure

Experimental protocol

For our six experimental sessions in Cologne, subjects were recruited via the online recruiting system ORSEE (Greiner, 2015) as well as via advertisements through the Faculty of Medicine at the University of Cologne. The recruiting process of subjects via ORSEE was the following: Students who registered in ORSEE for laboratory experiments at the CLER were invited via automatically generated e-mails to participate in the experimental conditions of our

experiment. Four further sessions were conducted in a seminar room of the medical school using the mobile labs of the CLER. Medical students were invited via e-mails sent from the Student Deanery of the Medical Faculty. In order to participate, they had to sign up for one of the sessions online in advance. When signing up, subjects did not know about the decision task, the composition of subjects, or about our research objective. This procedure guaranteed the random allocation of students to experimental treatments and excluded self-selection into treatments.

The recruiting process for subjects who participated in Bonn was the same. For a detailed experimental protocol for the sample of Bonn, please refer to Hennig-Schmidt et al. (2019).

The procedure in all treatments and both locations was as follows: Upon arrival, subjects were randomly allocated to cubicles. They were then given ample time to read the instructions and to ask any clarifying questions, which were answered in private. To verify that the subjects had understood the decision task, they had to answer a set of control questions. The experiment did not start unless all subjects had answered all control questions correctly. For the control questions, see Appendix A.1.5. Then, subjects anonymously decided, for six early-born infants, which birth weight to enter into each early-born's medical record after having seen the true birth weight on the computer screen. The introduction stressed that the subjects' decisions were confidential and anonymous and that the experimenter could not identify who had entered which birth weight. Subjects were free to leave the study at any time.

Before entering the birth weight of a specific baby, subjects could learn about the respective lump-sum remuneration, the treatment cost for each birth weight, and the resulting payment by clicking an information button on their computer screen. In the audit-and-fine treatments, i.e., 10ANF, 10AF, and 75AF, subjects could also learn about the lower bound of the preterm infant's weight on the day after birth. All monetary amounts were given in Taler, our experimental currency, the exchange rate being $1 \text{ Taler} = 0.01 \text{ EUR}$.⁶⁸ Subjects were also informed about the fixed payment F amounting to 400 Taler.

After having made all six decisions, the subjects were asked to answer a questionnaire

⁶⁸The exchange rate was calculated to result in an hourly wage of a student assistant at the Universities of Bonn and Cologne, which is about EUR 10.

on gender, age, integrity, personality traits and risk attitudes. Measures on personality comprise the Big-Five Inventory on personality traits by Rammstedt and John (2007) and the 18-items Integrity Scale by Schlenker (2008). They received EUR 4 for answering the questionnaire. Finally, they were paid by a confidential payment procedure that did not allow us to trace any individual subjects' decisions, and they subsequently left the laboratory or seminar room. Sessions lasted about 45 minutes, including filling in the post-experimental questionnaire. Subjects earned EUR 10.30 on average, including the payment of EUR 4 for answering the questionnaires.

Control questions

Control questions translated from German. The first nine questions apply to all treatment groups. Question 10 only applies to treatment groups including an audit. *Correct answers are written in italics.* [The text in brackets is for the conditions 10ANF, 10AF, and 75AF.]

Exercise Questions

Question 1

Are the following statements true or false?

- a) Each participant in the experiment decides which birth weight (in g) to write into the birth record for six early-born infants.

☒ True

☐ False. Correct answer: _____

- b) The birth weights displayed can have values between 1,400g and 1,500g.

☐ True

☒ False. Correct answer: between 1,150 g and 1,550 g.

- c) Birth weights can be recorded in 25-gram intervals.

☐ True

☒ False. Correct answer: in 50-gram intervals.

Question 2

Complete the following sentences.

- a) Each birth weight appears once.
- b) Your recorded birth weight determines which lump-sum reimbursement (in Taler) will be paid for the treatment of the early-born infant.

Question 3

Complete the following table by filling in the missing case-based lump sums in Taler.

Birth weight (in g)	1,150	1,200	1,250	1,300	1,350	1,400	1,450	1,500	1,550
Lump-sum reimbursement (in Taler)	380	<u>380</u>	200	<u>200</u>	<u>200</u>	200	200	<u>120</u>	120

Question 4

- a) How is your payment for each recorded birth weight determined?

Lump-sum reimbursement per case according to the recorded birth weight

minus

Costs for treatment of the early-born infant according to the birth weight displayed on the scale.

- b) Additionally, you receive a fixed amount of 400 Taler.

Question 5

Suppose that the scale has shown a birth weight of 1,250g and a weight of 1,200g has been recorded in the birth record.

- a) Which lump-sum reimbursement per case is paid out? 380 Taler
- b) How high are the arising costs? 260 Taler
- c) What is the payment? 120 Taler

Question 6

Suppose that the scale has shown a birth weight of 1,300g and a weight of 1,300g has been recorded in the birth record.

- a) Which lump-sum reimbursement per case is paid out? 200 Taler
- b) How high are the arising costs? 230 Taler
- c) What is the payment? -30 Taler

Question 7

Suppose that the scale has shown a birth weight of 1,400g and a weight of 1,300g has been recorded in the birth record.

- a) Which lump-sum reimbursement per case is paid out? 200 Taler
- b) How high are the arising costs? 180 Taler
- c) What is the payment? 20 Taler

Question 8

Suppose that the scale has shown a birth weight of 1,400g and a weight of 1,450g has been recorded in the birth record.

- a) Which lump-sum reimbursement per case is paid out? 200 Taler
- b) How high are the arising costs? 180 Taler
- c) What is the payment? 20 Taler

Question 9

Suppose that the scale has shown a birth weight of 1,500g and a weight of 1,450g has been recorded in the birth record.

- a) Which lump-sum reimbursement per case is paid out? 200 Taler

- b) How high are the arising costs? 140 Taler
- c) What is the payment? 60 Taler

Question 10

Please cross off what is not applicable:

- a) The newborn will be ~~washed~~ / *weighed* / ~~fed~~ again the day after birth. This weight will also be reported in the birth report, by a different person. It should be noted that a newborn infant loses ~~up to 5 gram~~ / *up to 5 percent* of its birth weight in the first 24 hours.
- b) An audit occurs ~~before~~ / ~~during~~ / *after* all birth weights have been recorded with a probability of ~~1%~~ / [10ANF, 10AF: 10%] / [75AF: 75%] / ~~90%~~ / ~~100%~~.
- c) If the weight measured on the day after birth is *higher* / ~~lower~~ than the *recorded* / ~~measured~~ birth weight, the health insurer assumes that the recorded birth weight was too low. This constitutes ~~an unfortunate error~~ / ~~an irrelevant fake entry~~ / *a serious fake entry*.
- d) In the event of the insurance provider's audit showing that the weight on the day after birth for *at least one* / ~~at least two~~ / ~~all~~ newborn is larger than the recorded weight, this constitutes a serious fake entry. This has [10AF, 75AF: an influence] / [10ANF: no influence] on your payment].

A.1.6 Integrity scale

Table A.1.3: The 18-items Integrity Scale by Schlenker (2008): Items

Item Number	Item wording
1	It is foolish to tell the truth when big profits can be made by lying. (R)
2	No matter how much money one makes, life is unsatisfactory without a strong sense of duty and character.
3	Regardless of concerns about principles, in today's world you have to be practical, adapt to opportunities, and do what is most advantageous for you. (R)
4	Being inflexible and refusing to compromise are good if it means standing up for what is right.
5	The reason it is important to tell the truth is because of what others will do to you if you don't, not because of any issue of right and wrong. (R)
6	The true test of character is a willingness to stand by one's principles, no matter what price one has to pay.
7	There are no principles worth dying for. (R)
8	It is important to me to feel that I have not compromised my principles.
9	If one believes something is right, one must stand by it, even if it means losing friends or missing out on profitable opportunities.
10	Compromising one's principles is always wrong, regardless of the circumstances or the amount that can be personally gained.
11	Universal ethical principles exist and should be applied under all circumstances, with no exceptions.
12	Lying is sometimes necessary to accomplish important, worthwhile goals. (R)
13	Integrity is more important than financial gain.
14	It is important to fulfill one's obligations at all times, even when nobody will know if one doesn't.
15	If done for the right reasons, even lying or cheating are ok. (R)
16	Some actions are wrong no matter what the consequences or justification.
17	One's principles should not be compromised regardless of the possible gain.
18	Some transgressions are wrong and cannot be legitimately justified or defended regardless of how much one tries.

Notes. Subjects are asked to indicate the extent of their agreement or disagreement where 1 = strongly disagree, 2 = disagree, 3 = neither disagree nor agree, 4 = agree, and 5 = strongly agree. Items marked (R) are reverse scored.

A.2 Additional analyses

A.2.1 Comparison of samples

In order to test whether student samples in Cologne in Bonn differ, we first compare sample characteristics and second upcoding behavior for treatments NANF and 10AF which were conducted in both locations. In total, 23 subjects participated in Cologne and 98 in Bonn. Subsamples do not differ either by sample characteristics (see Table A.2.1) nor by upcoding behavior.

Table A.2.1: Comparison of sample characteristics between Bonn and Cologne

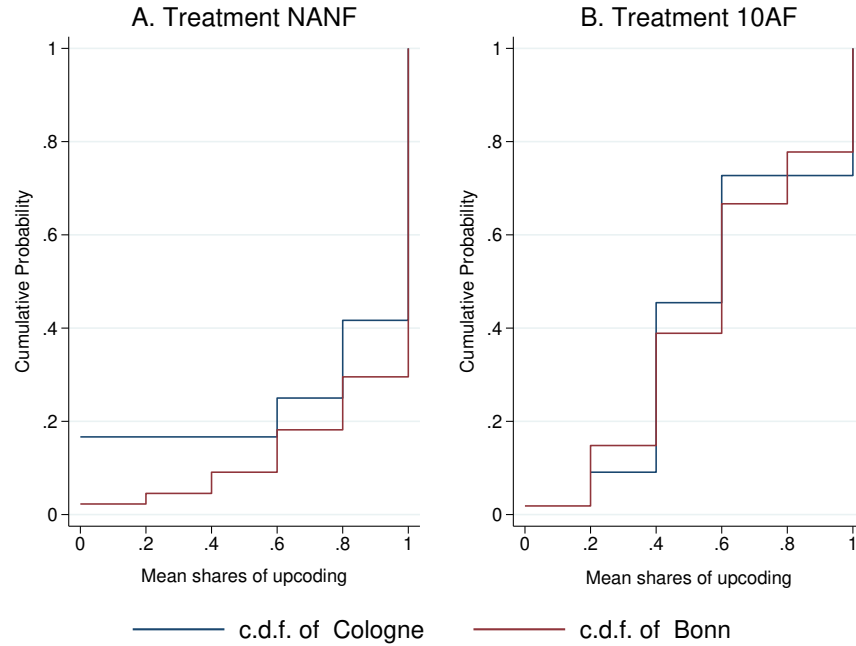
	Samples		<i>p</i> -value
	Bonn	Cologne	
Age (Mean, s.d.)	23.1 (5.17)	22.9 (2.66)	0.873
Share of females	46.9%	34.8%	0.291
Share med. students	52.0%	34.8%	0.136
<i>N</i> subjects	98	23	

Notes. This table compares the sample characteristics for both locations. The last column reports *p*-values of non-parametric analyses for differences. We perform a χ^2 -test for differences in shares by gender and medical studies and a t-test for age differences.

To test whether upcoding behavior differs between student samples in Cologne and Bonn, we compare the distributions of upcoding behavior for both locations. To this end, we analyze whether the mean shares of upcoding per subject, calculated as the sum of upcoded birth weight entries over the overall number of decision in which upcoding is possible ($i = 5$), differ between Bonn and Cologne for the treatments which were performed in both locations, i.e., treatments NANF and 10AF. Panel A of Figure A.2.1 shows the cumulative distribution functions (cdf) of mean shares per subject for the baseline treatment (NANF), and Panel B for our 10%-audit-and-fine treatment (10AF). The blue lines refer to the cdfs of mean shares of upcoding behavior per subject for Cologne, while the red lines refer to the same for Bonn. The similar course of the lines indicates that upcoding behavior between both locations does not differ a lot. Moreover, we find no statistical support for a difference in upcoding behavior between locations. First, employing the Kolmogorov-Smirnov test for equality of distribution functions, we cannot reject the null hypotheses of identical distributions

for Bonn and Cologne ($p > 0.100$). Second, there is no statistical difference either between the empirical characteristic functions of both locations based on the Epps-Singleton test ($p > 0.100$). Since both tests result in p -values above significance level, there is no evidence of a difference between the distributions by location. Hence, we combine the data sets of Bonn and Cologne for the NANF and the 10AF treatment.

Figure A.2.1: Shares of upcoding per subject for treatments NANF and 10AF, differentiated by location



Notes. This figure shows the cumulative distribution functions (c.d.f.) of mean shares of upcoding per subject differentiated by location for the baseline treatment (NANF; Panel A on the left side) and 10%-audit-and-fine treatment (10AF; Panel B on the right side). The blue line refers to subjects of Cologne and the red to subjects of Bonn. In NANF, 44 subjects participated in Bonn, and 12 in Cologne; in 10AF, 54 subjects participated in Bonn, and 11 in Cologne.

A.2.2 Descriptive statistics on individual characteristics, differentiated by treatment

Table A.2.2: Individual characteristics by treatment

	A. Overall		B. NANF		C. 10ANF		D. 10AF		E. 75AF	
	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.
Age	23.57	5.20	22.68	2.32	24.84	6.35	23.38	6.17	23.92	5.11
Share of females (in %)	44.7		35.7		44.7		52.3		44.7	
<i>Personality traits</i>										
Extraversion	0.19	0.42	0.16	0.30	0.22	0.55	0.23	0.37	0.12	0.51
Agreeableness	0.29	0.30	0.32	0.26	0.25	0.34	0.30	0.31	0.24	0.30
Conscientiousness	0.33	0.37	0.25	0.30	0.39	0.41	0.32	0.40	0.43	0.37
Neuroticism	-0.03	0.44	0.03	0.34	-0.02	0.52	-0.04	0.42	-0.13	0.53
Openness	0.31	0.46	0.20	0.36	0.38	0.60	0.34	0.39	0.38	0.51
Integrity	0.14	0.17	0.15	0.14	0.10	0.23	0.14	0.17	0.17	0.15
N	197		56		38		65		38	

Notes. This table shows the share of females as well as means and standard deviations for individual characteristics (i.e., age, personality traits (Big-Five Inventory), integrity (Schlenker, 2008), by treatments. All personality characteristics such as personality traits and integrity are measured on a scale from 1 to +1, with -1= theoretical minimum, 0= neutral midpoint, and 1= theoretical maximum.

A.2.3 Additional analyses on treatment effects

Table A.2.3: Differences of birth weight reporting by treatments, proportion of participants (in %)

	Upcoding		Honest	
	Detectable	Undetectable	reporting	Unclassified
NANF	62.1	20.0	11.4	6.4
10ANF	52.1	24.2	14.2	9.5
10AF	30.5	27.7	33.5	8.3
75AF	7.9	28.9	54.2	8.9

Notes. This table shows the proportions in reporting types of participants by treatments. Upcoding is defined as payment-increasing misreports of birth weights. Detectable upcoding reflects upcoding which is detectable in case of an audit (upcoding by 100g or more). Undetectable upcoding means that individuals misreport birth weights by only 50g, which cannot be detected even in case of an audit. Honest reporting considers honest birth weight reporting. We include all true birth weights except of 1,200g where payment-increasing upcoding is not possible ($n=197$ subjects, $k=985$ decisions). All birth weights reports which cannot be classified into the above defined categories are considered as unclassified. The proportions of unclassified observations do not differ significantly with treatments.

Table A.2.4: Descriptive statistics and analyses of differences in upcoding between treatments (proportion of participants)

A. Honest reporting when only detectable upcoding is possible (at 1,300g, 1,350g, and 1,400g; $k=591$ decisions)			
Detection prob.	No fine	Fine	Δ , in pp (p -value)
0% (no audit)	14.9% (NANF)	—	
10%	21.1% (10ANF)	48.7% (10AF)	27.7 (<0.001)
75%	—	75.4% (75AF)	
Δ , in pp (p -value)	6.2 (0.357)	26.7 (0.001)	
B. Honest reporting when undetectable upcoding is possible (at 1,250g; $k=197$ decisions)			
Detection prob.	No fine	Fine	Δ , in pp (p -value)
0% (no audit)	7.1% (NANF)	—	
10%	2.6% (10ANF)	6.2% (10AF)	3.5 (0.374)
75%	—	23.7% (75AF)	
Δ , in pp (p -value)	-4.5 (0.297)	17.5 (0.020)	
C. Honest reporting when undetectable and detectable upcoding are feasible (at 1,500g; $k=196$ decisions)			
Detection prob.	No fine	Fine	Δ , in pp (p -value)
0% (no audit)	5.5% (NANF)	—	
10%	5.3% (10ANF)	15.4% (10AF)	10.1 (0.079)
75%	—	21.1% (75AF)	
Δ , in pp (p -value)	-0.2 (0.968)	5.7 (0.478)	

Notes. This table reports the proportions of honest reporting by treatments and differences in proportions. Panel A refers to proportion of honest reporting for infants for whom only detectable upcoding is possible (1,300g, 1,350g, and 1,400g; $k=591$ decisions). Upcoding is defined as payment-increasing misreports of birth weights. Panel B refers to infants for whom undetectable upcoding is possible (1,250g, $k=197$ decisions). Undetectable upcoding means that individuals misreport birth weights by only 50g, which cannot be detected by an audit. Panel C refers to infants for whom undetectable upcoding is possible but less gainful than detectable upcoding (1,500g, $k=196$ decisions). Treatment effect differences Δ are based on the marginal effects based on logit models reported A.2.6 and multinomial logit models for upcoding at 1,500g reported in Table A.2.7 of the Appendix.

Table A.2.5: Descriptive statistics and analyses of differences in detectable and undetectable upcoding at 1,500g (proportions of participants)

A. Detectable upcoding (at 1,500g; $k=196$ decisions)			
Detection prob.	No fine	Fine	Δ , in pp (p -value)
0% (no audit)	72.7% (NANF)	—	
10%	55.3% (10ANF)	30.8% (10AF)	−24.5 (0.013)
75%	—	7.9% (75AF)	
Δ , in pp (p -value)	−17.5 (0.082)	−22.9 (0.001)	
B. Undetectable upcoding (at 1,500g; $k=196$ decisions)			
Detection prob.	No fine	Fine	Δ , in pp (p -value)
0% (no audit)	21.8% (NANF)	—	
10%	39.5% (10ANF)	53.9% (10AF)	14.4 (0.153)
75%	—	71.1% (75AF)	
Δ , in pp (p -value)	17.7 (0.068)	17.2 (0.073)	

Notes. This table reports the proportions of upcoding cases at 1,500g by treatments and differences in proportions ($k=196$ decisions). "Detectable upcoding" takes place when an individual reports a weight of 1,250g or lower, "Undetectable upcoding" takes place when an individual reports a (fraudulent birth) weight of 1,450g. Estimates for "Honest reporting" are reported in Table 1.2. We excluded one subject with a birth-weight entry of 1,550g. Treatment effects (Δ) are differences in marginal effects (percentage points, pp). Full regression results are reported in Table A.2.7 of the Appendix.

Table A.2.6: Treatment effects on the likelihood of detectable upcoding, undetectable upcoding and honest reporting, logit regression models, MEMs

	Honest reporting			
	A. At 1,300g, 1,350g and 1,400g		B. At 1,250g	
	(1)	(2)	(3)	(4)
Base cat.: NANF (No-audit-no-fine)				
10ANF (10%-audit-no-fine)	0.062 (0.067)	0.074 (0.064)	−0.045 (0.043)	−0.056 (0.043)
10AF (10%-audit-and-fine)	0.338*** (0.065)	0.400*** (0.065)	−0.010 (0.046)	−0.017 (0.046)
75AF (75%-audit-and-fine)	0.606*** (0.071)	0.615*** (0.075)	0.165** (0.077)	0.143* (0.080)
Effect differences				
Δ (10ANF, 10AF)	0.277*** (0.074)	0.326*** (0.079)	0.035 (0.040)	0.039 (0.038)
Δ (10ANF, 75AF)	0.544*** (0.079)	0.541*** (0.088)	0.211*** (0.074)	0.199*** (0.075)
Δ (10AF, 75AF)	0.267*** (0.077)	0.214** (0.088)	0.175** (0.075)	0.160* (0.088)
Birth-weight controls	No	Yes	No	No
Individual controls	No	Yes	No	Yes
k decisions	591	591	197	197
n subjects	197	197	197	197
Pseudo R^2	0.170	0.214	0.089	0.120

Notes. This table reports marginal effects at the means of the covariates (MEMs) based on logit models. Standard errors are shown in parentheses. For Panel A standard errors are clustered at the individual subject level. Honest reporting is defined as reporting the true birth weights. Upcoding is defined as payment-increasing misreports of birth weights. Models (1) and (2) are estimated for true birth weights, for which only detectable upcoding is possible (1,300g, 1,350g, and 1,400g; $k=591$ decisions). Models (3) and (4) are estimated for true birth weight of 1,250g ($k=197$ decisions), for which undetectable upcoding is possible. Undetectable upcoding means that individuals misreport birth weights by only 50g, which cannot be detected by an audit. Individual control variables comprise gender, age, medical major, personality traits (Big-Five Inventory), and integrity (Schlenker, 2008). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2.7: Treatment effects on the likelihood of detectable upcoding, undetectable upcoding and honest reporting at true birth weight 1,500g, multinomial logit regression, MEMs

	A. Without individual controls			B. With individual controls		
	Detectable upcoding (1)	Undetectable upcoding (2)	Honest reporting (3)	Detectable upcoding (4)	Undetectable upcoding (5)	Honest reporting (6)
Base cat.: NANF (No-audit-no-fine)						
10ANF (10%-audit-no-fine)	-0.175* (0.101)	0.177* (0.097)	-0.002 (0.047)	-0.184* (0.108)	0.192* (0.103)	-0.008 (0.038)
10AF (10%-audit-and-fine)	-0.420*** (0.083)	0.320*** (0.083)	0.099* (0.054)	-0.513*** (0.087)	0.394*** (0.088)	0.120** (0.056)
75AF (75%-audit-and-fine)	-0.648*** (0.074)	0.492*** (0.092)	0.156** (0.073)	-0.685*** (0.079)	0.553*** (0.095)	0.131* (0.070)
Effect differences						
Δ (10ANF, 10AF)	-0.245** (0.099)	0.144 (0.101)	0.101* (0.058)	-0.329*** (0.108)	0.202* (0.109)	0.128** (0.056)
Δ (10ANF, 75AF)	-0.474*** (0.092)	0.316*** (0.108)	0.158** (0.075)	-0.500*** (0.098)	0.361*** (0.112)	0.139** (0.069)
Δ (10AF, 75AF)	-0.229*** (0.072)	0.172* (0.096)	0.057 (0.080)	-0.171** (0.075)	0.160 (0.101)	0.012 (0.082)
k decisions	196	196	196	196	196	196
n subjects	196	196	196	196	196	196

Notes. This table reports marginal effects at the means of the covariates (MEMs) based on multinomial logit models. Standard errors are shown in parentheses. All models consider reporting behavior for a true birth weight of 1,500g ($k=196$ decisions). With respect to the entered birth weight, a subject's reporting behavior is classified into one of three categories. "Detectable upcoding" takes place when an individual reports a (fraudulent birth) weight lower than 1,450g, "Undetectable upcoding" takes place when an individual reports a (fraudulent birth) weight of 1,450g. "Honest reporting" takes place when an individual reports the true birth weight of 1,500g. We excluded one subject with a birth-weight entry of 1,550g. In Panel B, we control for individual characteristics, i.e., gender, age, medical major, personality traits (Big-Five Inventory), and integrity (Schlenker, 2008). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.2.4 Distribution of individuals' decisions

Table A.2.8: Frequency of reported birth weights by true birth weights: No-audit-no-fine treatment

True weight	Reported weight									Total
	1150	1200	1250	1300	1350	1400	1450	1500	1550	
1200 (abs.)	5	46	2	1	0	0	1	1	0	56
1200 (%)	9	82	4	2	0	0	2	2	0	100
1250	7	44	4	0	0	1	0	0	0	56
	13	79	7	0	0	2	0	0	0	100
1300	8	40	2	6	0	0	0	0	0	56
	14	71	4	11	0	0	0	0	0	100
1350	10	34	0	1	10	0	0	0	1	56
	18	61	0	2	18	0	0	0	2	100
1400	8	35	0	1	3	9	0	0	0	56
	14	63	0	2	5	16	0	0	0	100
1500	9	30	0	0	0	1	12	3	1	56
	16	54	0	0	0	2	21	5	2	100
Total	47	229	8	9	13	11	13	4	2	336
	14	68	2	3	4	3	4	1	1	100

Notes. Light gray cells indicate understated birth weights. Dark gray cells show honest reporting of birth weights. White cells denote overstated birth weights. The numbers in the first (second) line for each birth weight indicate absolute frequencies (percentages). Reimbursement thresholds are marked by vertical lines. The data set for the no-audit-no-fine treatment consists of 336 entries ($n = 56$ subjects).

Table A.2.9: Frequency of reported birth weights by true birth weights: 10%-audit-no-fine treatment

True weight	Reported weight									Total
	1150	1200	1250	1300	1350	1400	1450	1500	1550	
1200 (abs.)	9	29	0	0	0	0	0	0	0	38
1200 (%)	24	76	0	0	0	0	0	0	0	100
1250	5	31	1	1	0	0	0	0	0	38
	13	82	3	3	0	0	0	0	0	100
1300	8	22	2	5	1	0	0	0	0	38
	21	58	5	13	3	0	0	0	0	100
1350	7	17	0	3	10	1	0	0	0	38
	18	45	0	8	26	3	0	0	0	100
1400	5	19	0	0	2	9	3	0	0	38
	13	50	0	0	5	24	8	0	0	100
1500	6	15	0	0	0	0	15	2	0	38
	16	39	0	0	0	0	39	5	0	100
Total	40	133	3	9	13	10	18	2	0	228
	18	58	1	4	6	4	8	1	0	100

Notes. Light (Medium) gray cells indicate understated birth weight reports which are detectable (undetectable) in case of an audit. Dark gray cells show honest reporting of birth weights. White cells denote overstated birth weights. Numbers in the first (second) line for each true weight indicate absolute frequencies (percentages). Reimbursement thresholds are marked by vertical lines. Data set for the 10% and no fine treatment consists of 228 entries ($n = 38$ subjects).

Table A.2.10: Frequency of reported birth weights by true birth weights: 10%-audit-and-fine treatment

True weight	Reported weight									Total
	1150	1200	1250	1300	1350	1400	1450	1500	1550	
1200 (abs.)	5	59	0	0	0	0	0	0	1	65
1200 (%)	8	91	0	0	0	0	0	0	2	100
1250	6	55	4	0	0	0	0	0	0	65
	9	85	6	0	0	0	0	0	0	100
1300	7	28	5	23	2	0	0	0	0	65
	11	43	8	35	3	0	0	0	0	100
1350	7	15	0	5	35	2	1	0	0	65
	11	23	0	8	54	3	2	0	0	100
1400	6	16	0	0	3	37	3	0	0	65
	9	25	0	0	5	57	5	0	0	100
1500	7	13	0	0	0	0	35	10	0	65
	11	20	0	0	0	0	54	15	0	100
Total	38	186	9	28	40	39	39	10	1	390
	10	48	2	7	10	10	10	3	0	100

Notes. Light (Medium) gray cells indicate understated birth weight reports which are detectable (undetectable) in case of an audit. Dark gray cells show honest reporting of birth weights. White cells denote overstated birth weights. The numbers in the first (second) line for each true weight indicate absolute frequencies (percentages). Reimbursement thresholds are marked by vertical lines. The data set for the 10% and fine treatment consists of 390 entries ($n = 65$ subjects).

Table A.2.11: Frequency of reported birth weights by true birth weights: 75%-audit-and-fine treatment

True weight	Reported weight									Total
	1150	1200	1250	1300	1350	1400	1450	1500	1550	
1200 (abs.)	7	31	0	0	0	0	0	0	1	38
1200 (%)	18	82	0	0	0	0	0	0	2	100
1250	1	28	9	0	0	0	0	0	0	38
	3	74	24	0	0	0	0	0	0	100
1300	3	2	4	28	2	0	0	0	0	38
	8	5	11	74	3	0	0	0	0	100
1350	1	2	0	5	28	2	0	0	0	38
	3	5	0	13	74	5	0	0	0	100
1400	2	2	0	0	3	30	1	0	0	38
	5	5	0	0	8	79	3	0	0	100
1500	1	2	0	0	0	0	27	8	0	38
	3	5	0	0	0	0	71	21	0	100
Total	15	67	13	33	32	32	28	8	0	228
	7	29	6	14	14	14	12	4	0	100

Notes. Light (Medium) gray cells indicate understated birth weight reports which are detectable (undetectable) in case of an audit. Dark gray cells show honest reporting of birth weights. White cells denote overstated birth weights. The numbers in the first (second) line for each birth weight indicate absolute frequencies (percentages). Reimbursement thresholds are marked by vertical lines. The data set for the 75% and fine treatment consists of 228 entries ($n = 38$ subjects).

B Appendix to Chapter 2

B.1 Additional information about the experiment

B.1.1 Laboratory setup

Figure B.1.1: Mobile and computer laboratory



Notes. This figure shows the laboratory setup for the computer laboratory experiments at elfe at the University of Duisburg-Essen (left panel) for our student sample and the mobile laboratory setup of elfe at the Academy for Training and Education in Bad Nauheim for our physician sample.

B.1.2 Sample

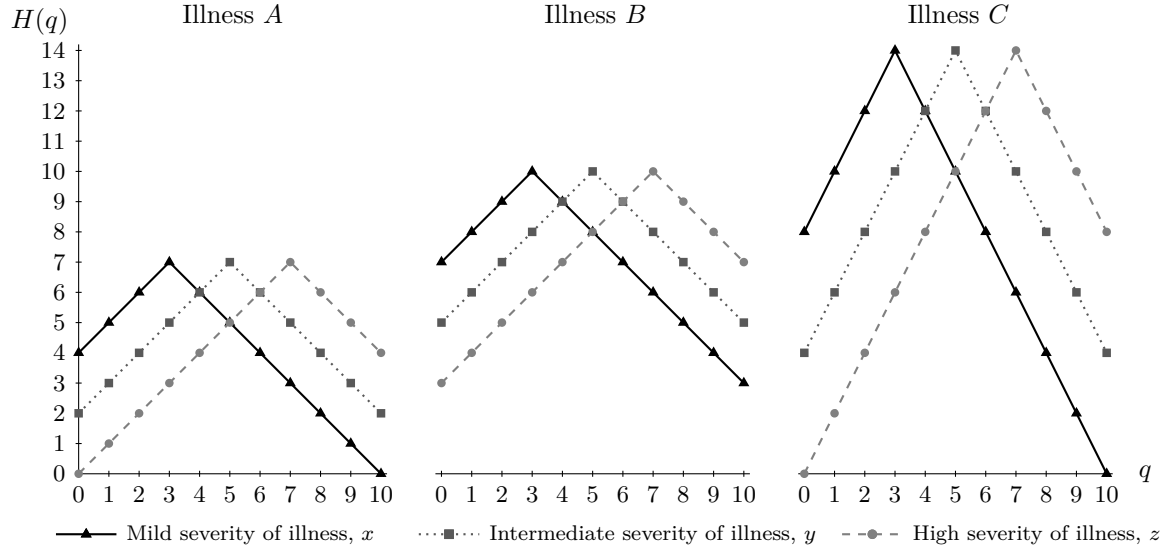
Table B.1.1: Sample characteristics

	Sample			
	Full sample	Medical students	Non-medical students	Physicians
A. All payment systems				
<i>Main characteristics</i>				
Male	40.2%	27.3%	53.5%	40.0%
Age (Mean, s.d.)	28.6 (9.8)	25.2 (7.0)	24.3 (3.4)	45.3 (6.7)
<i>Personality traits</i> (Mean, s.d.)				
Extraversion	3.6 (0.83)	3.7 (0.84)	3.5 (0.90)	3.5 (0.56)
Neuroticism	2.8 (0.97)	2.6 (0.91)	3.0 (0.92)	2.6 (1.10)
Openness	3.6 (0.92)	3.8 (0.88)	3.3 (0.97)	3.6 (0.82)
Conscientiousness	3.6 (0.81)	3.6 (0.69)	3.2 (0.81)	4.3 (0.53)
Agreeableness	3.1 (0.71)	3.3 (0.63)	2.8 (0.73)	3.1 (0.64)
<i>N</i>	107	44	43	20
B. FFS				
<i>Main characteristics</i>				
Male	44.2%	40.9%	50.0%	40.0%
Age (Mean, s.d.)	28.4 (9.9)	24.3 (4.5)	24.1 (4.0)	45.9 (7.2)
<i>Personality traits</i> (Mean, s.d.)				
Extraversion	3.7 (0.86)	3.8 (0.89)	3.7 (0.92)	3.4 (0.66)
Neuroticism	2.8 (0.92)	2.4 (0.82)	3.1 (0.85)	3.0 (0.08)
Openness	3.5 (0.93)	3.9 (0.85)	3.3 (1.00)	3.5 (0.80)
Conscientiousness	3.5 (0.79)	3.5 (0.66)	3.1 (0.82)	4.1 (0.61)
Agreeableness	3.1 (0.67)	3.3 (0.66)	2.9 (0.61)	3.2 (0.63)
<i>N</i>	52	22	20	10
C. CAP				
<i>Main characteristics</i>				
Male	36.4%	13.6%	56.5%	40.0%
Age (Mean, s.d.)	28.8 (9.9)	26.1 (8.8)	24.6 (2.9)	44.6 (8.2)
<i>Personality traits</i> (Mean, s.d.)				
Extraversion	3.4 (0.79)	3.5 (0.79)	3.4 (0.88)	3.4 (0.58)
Neuroticism	2.7 (1.02)	2.8 (0.99)	2.9 (0.99)	2.3 (1.11)
Openness	3.6 (0.91)	3.7 (0.92)	3.4 (0.95)	4.0 (0.69)
Conscientiousness	3.7 (0.82)	3.8 (0.70)	3.3 (0.81)	4.5 (0.44)
Agreeableness	3.0 (0.76)	3.2 (0.61)	2.9 (0.83)	3.0 (0.90)
<i>N</i>	55	22	23	10

Notes. This table presents summary statistics of subjects' characteristics for (i) the full sample of our experiment, (ii) for medical and (iii) non-medical students in the laboratory experiment and (iv) for physicians in the artifactual field experiment. We further differentiate between payment systems.

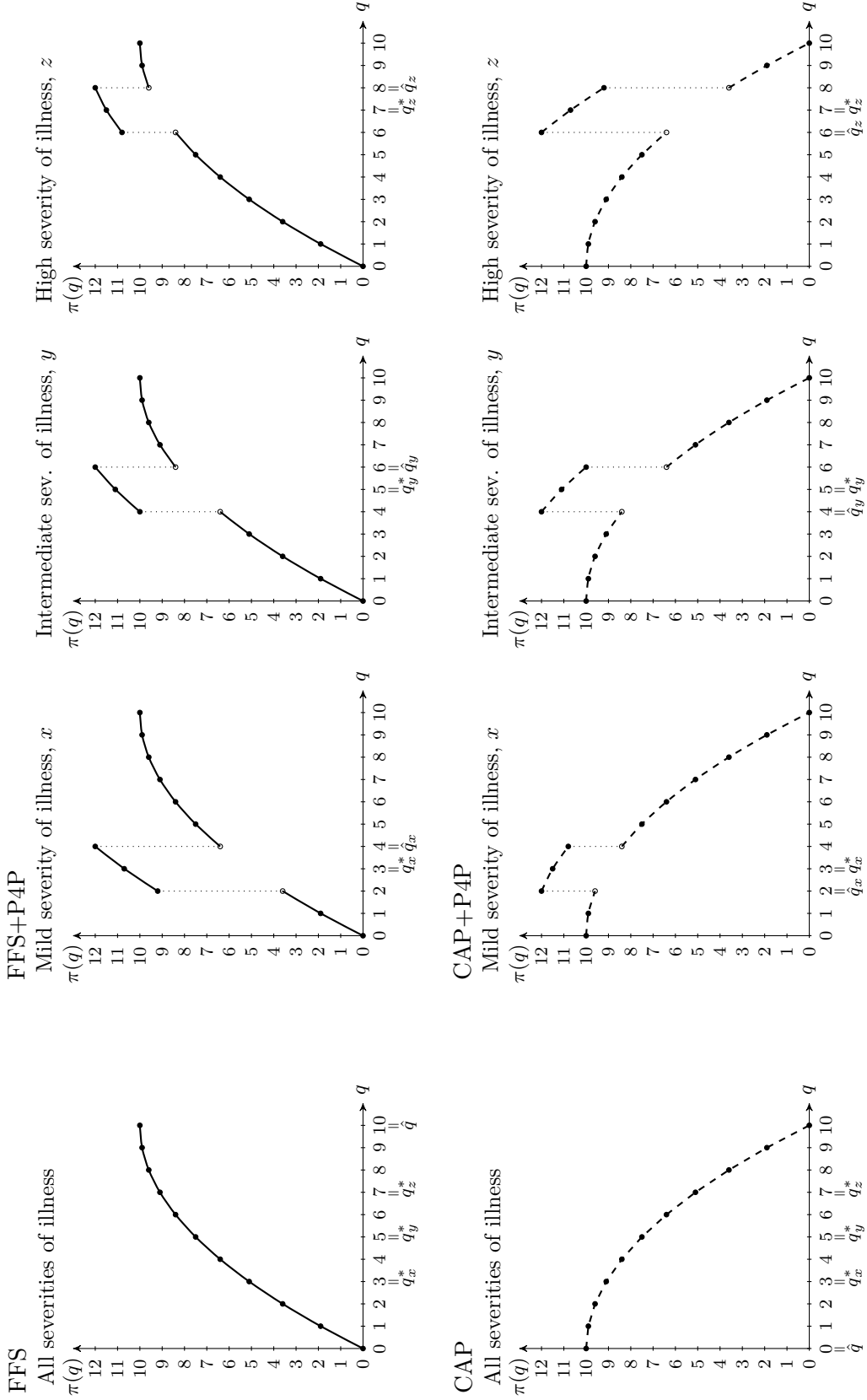
B.1.3 Parameters of the experiment

Figure B.1.2: Patient health benefits by illness and severity of illness



Notes. This figure illustrates patient health benefit parameters $H(q)$ for illnesses $k = A, B, C$ and severities of illness $l = x, y, z$ on the quantity interval from 0 to 10. The left panel shows patient benefits for illness A, the middle panel for illness B, and the right panel for illness C. The black solid line indicates severity of illness x , the grey dotted line severity of illness y , and the grey dashed line severity of illness z . For illness A and B, $\theta = 1$ and for illness C, $\theta = 2$. Notice that the patient health benefits are kept constant for all payment conditions.

Figure B.1.3: Profit parameters in FFS/FFS+P4P and CAP/CAP+P4P



Notes. The upper panel of the figure illustrates profits in FFS and FFS+P4P, the lower panel analogously illustrates profits in CAP and CAP+P4P. Under basic payments, profits increase (in FFS) and decrease (in CAP) continuously, regardless of the severity of illness on the quantity interval. In the pay for performance conditions, a bonus payment is granted if the performance threshold $|q - q^*| \leq 1$ is reached. As the patient-optimal quantity q^* depends on the severity of illness, the performance thresholds differ accordingly. In the basic payment condition, the profit-maximizing quantity is $\hat{q}=10$ in FFS and $\hat{q}=0$ in CAP, respectively. In the pay for performance condition, \hat{q} changes depending on the severity of illness.

Table B.1.2: Parameters of main experimental conditions

	Quantity (q)										
	0	1	2	3	4	5	6	7	8	9	10
Patient benefit											
B_{Ax}	4	5	6	7	6	5	4	3	2	1	0
B_{Ay}	2	3	4	5	6	7	6	5	4	3	2
B_{Az}	0	1	2	3	4	5	6	7	6	5	4
B_{Bx}	7	8	9	10	9	8	7	6	5	4	3
B_{By}	5	6	7	8	9	10	9	8	7	6	5
B_{Bz}	3	4	5	6	7	8	9	10	9	8	7
B_{Cx}	8	10	12	14	12	10	8	6	4	2	0
B_{Cy}	4	6	8	10	12	14	12	10	8	6	4
B_{Cz}	0	2	4	6	8	10	12	14	12	10	8
Costs											
c	0.0	0.1	0.4	0.9	1.6	2.5	3.6	4.9	6.4	8.1	10.0
FFS											
p	0.0	2.0	4.0	6.0	8.0	10.0	12.0	14.0	16.0	18.0	20.0
π	0.0	1.9	3.6	5.1	6.4	7.5	8.4	9.1	9.6	9.9	10.0
CAP											
L	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
π	10.0	9.9	9.6	9.1	8.4	7.5	6.4	5.1	3.6	1.9	0.0
FFS+P4P											
p	0.0	2.0	4.0	6.0	8.0	10.0	12.0	14.0	16.0	18.0	20.0
b_x	0.0	0.0	5.6	5.6	5.6	0.0	0.0	0.0	0.0	0.0	0.0
b_y	0.0	0.0	0.0	0.0	3.6	3.6	3.6	0.0	0.0	0.0	0.0
b_z	0.0	0.0	0.0	0.0	0.0	0.0	2.4	2.4	2.4	0.0	0.0
π_x	0.0	1.9	9.2	10.7	12.0	7.5	8.4	9.1	9.6	9.9	10.0
π_y	0.0	1.9	3.6	5.1	10.0	11.1	12.0	9.1	9.6	9.9	10.0
π_z	0.0	1.9	3.6	5.1	6.4	7.5	10.8	11.5	12.0	9.9	10.0
CAP+P4P											
L	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
b_x	0.0	0.0	2.4	2.4	2.4	0.0	0.0	0.0	0.0	0.0	0.0
b_y	0.0	0.0	0.0	0.0	3.6	3.6	3.6	0.0	0.0	0.0	0.0
b_z	0.0	0.0	0.0	0.0	0.0	0.0	5.6	5.6	5.6	0.0	0.0
π_x	10.0	9.9	12.0	11.5	10.8	7.5	6.4	5.1	3.6	1.9	0.0
π_y	10.0	9.9	9.6	9.1	12.0	11.1	10.0	5.1	3.6	1.9	0.0
π_z	10.0	9.9	9.6	9.1	8.4	7.5	12.0	10.7	9.2	1.9	0.0

Notes. This table shows the parameters used in our experiment for all payment conditions. p is the fee per service rendered to a patient in FFS, L is the lump-sum payment in CAP, b_i^\bullet is the bonus paid when the quality requirement is met in FFS+P4P (CAP+P4P), and π is the physician's profit.

B.1.4 Instructions of the experiment

Notice that the text in squared brackets denotes [Capitation, CAP] conditions.

Welcome to the Experiment!

You are participating in an economic experiment on decision behavior. You and the other participants will be asked to make decisions for which you can earn money. Your payoff depends on the decisions you make. At the end of the experiment, your payoff will be converted to Euro and paid to you in cash. During the experiment, all amounts are presented in the experimental currency Taler. 10 Taler equal 8 Euro. The experiment will take about 90 minutes and consists of two parts. You will receive detailed instructions before each part. Note that none of your decisions in either part have any influence on the other part of the experiment.

Part I of the experiment

Please read the instructions carefully. We will approach you in about five minutes to answer any questions you may have. If you have questions at any time during the experiment, please raise your hand and we will come to you. Part I of the experiment consists of 9 rounds of decision situations.

Decision situation

In each round, you are in the role of a physician and decide on medical treatment for a patient. That is, you determine the quantity of medical services you wish to provide to the patient for a given illness and a given severity of this illness. Each patient is characterized by one of three illnesses (A, B, C), each of which can occur in three different degrees of severity (x, y, z). In each consecutive decision round, you will face one patient who is characterized by one of the 9 possible combinations of illnesses and degrees of severity (in random order). Your decision is to provide each of these 9 patients with a quantity of 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 medical services.

Payment

In each round, you receive a fee-for-service [capitation] remuneration for treating the patient. Your remuneration increases with the amount of medical treatment [irrespective of the amount of medical treatment] you provide. You also incur costs for treating the patient, which likewise depend on the quantity of services you provide. Your profit for each decision is calculated by subtracting these costs from the fee-for-service [capitation] remuneration. Each quantity of medical service yields a particular benefit for the patient—contingent on his illness and severity. Hence, in choosing the medical services you provide, you determine not only your own profit but also the patient's benefit.

In each round you will receive detailed information on your screen (see below) for the respective patient, the illness, your amount of fee-for-service [capitation] remuneration—for each possible amount of medical treatment—your costs, profit, as well as the benefit for the patient with the corresponding illness and severity.

Payoff

At the end of the experiment, one of the 9 rounds in part *I* will be chosen at random. Your profit in that round will be paid to you in cash.

For this part of the experiment, no patients are physically present in the laboratory. Yet the patient benefit does accrue to a real patient: The amount resulting from your decision will be transferred to the Christoffel Blindenmission Deutschland e.V., 64625 Bensheim, which will use the money for enabling the treatment of patients with eye cataract.

The transfer of money to the Christoffel Blindenmission Deutschland e.V. will be carried out after the experiment by the experimenter and one participant. The participant completes a money transfer form, filling in the total patient benefit (in Euro) resulting from the decisions made by all participants in the randomly chosen situation. This form prompts the payment of the designated amount to the Christoffel Blindenmission Deutschland e.V. by the finance

Fee-for-service, FFS:

Patient 1 with illness

Quantity of medical treatment	Your fee-for-service payment (in Taler)	Your costs (in Taler)	Your profit (in Taler)	Benefit of the Patient with illness and severity (in Taler)

Which quantity of medical treatment do you want to provide?

Your decision:

[Capitation, CAP:]

Patient 1 with illness

Quantity of medical treatment	Your capitation payment (in Taler)	Your costs (in Taler)	Your profit (in Taler)	Benefit of the patient with illness and severity (in Taler)

Which quantity of medical treatment do you want to provide?

Your decision:

department of the University of Duisburg-Essen. The form is then sealed in a stamped envelope and deposited in the nearest mailbox by the participant and the experimenter.

After the entire experiment is completed, one participant is chosen at random to oversee the money transfer to the Christoffel Blindenmission Deutschland e.V. The participant receives an additional compensation of 5 Euro for this task. The participant certifies that the process has been completed as described here by signing a statement that can be inspected by all participants at the office of the Chair of Quantitative Economic Policy. A receipt of the bank transfer to the Christoffel Blindenmission Deutschland e.V. may also be viewed here.

Comprehension questions

Prior to the decision rounds, we kindly ask you to answer a few comprehension questions. They are intended to help you familiarize yourself with the decision situations. If you have any questions about this, please raise your hand. Part *I* of the experiment will begin once all participants have answered all comprehension questions correctly.

Part II of the experiment

Please read the instructions carefully. We will approach you in about five minutes to answer any questions you may have. If you have questions at any time during the experiment, please raise your hand and we will come to you. Part *II* of the experiment also consists of 9 rounds of decision situations.

Decision situation

As in part *I* of the experiment, you take on the role of a physician in each round and decide on medical treatment for a patient. That is, you determine the quantity of medical services you wish to provide to the patient for a given illness and a given severity of this illness.

Each patient is characterized by one of three illnesses (A, B, C), each of which can occur in three different degrees of severity (x, y, z). In each consecutive decision round, you will face one patient who is characterized by one of the 9 possible combinations of illnesses and degrees of severity (in random order). Your decision is to provide each of these 9 patients with a quantity of 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 medical services.

Payment

In each round, you are remunerated for treating the patient. In each round, you receive a fee-for-service [capitation] remuneration for treating the patient. Your remuneration increases with the amount of medical treatment [is irrespective of the amount of medical

treatment] you provide. In addition to this, in each round you receive a bonus payment, in case the quantity of medical services you provide is equal to the one that results in the highest benefit for the patient, or deviates by one quantity from the latter. You also incur costs for treating the patient, which likewise depend on the quantity of services you provide. Your profit for each decision is calculated by subtracting these costs from the sum of your fee-for-service [capitation] remuneration and bonus payment.

As in part *I*, every quantity of medical service yields a particular benefit for the patient contingent on his illness and severity. Hence, in choosing the medical services you provide, you determine not only your own profit, but also the patient's benefit.

In each round, you will receive detailed information on your screen (see below) for the respective patient, the illness, your amount of fee-for-service [capitation] remuneration—for each possible amount of medical treatment, the amount of your bonus payment, your costs, profit, as well as the benefit for the patient with the corresponding illness and severity.

Payoff

At the end of the experiment, one of the 9 rounds of part *II* will be chosen at random. Your profit in this round will be paid to you in cash, in addition to your payment from the round chosen for part *I* of the experiment. After the experiment is over, please remain seated until the experimenter asks you to step forward. You will receive your payment at the front of the laboratory before exiting the room.

As in part *I*, no patients are physically present in the laboratory for part *II* of the experiment. Yet the patient benefit does accrue to a real patient: The amount resulting from your decision will be transferred to the Christoffel Blindenmission Deutschland e.V., 64625 Bensheim, which will use the money for enabling the treatment of patients with eye cataract.

The process for transferring the money to the Christoffel Blindenmission Deutschland e.V., as described for part *I* of the experiment, will be carried out by the experimenter and one participant.

FFS+P4P:

Patient 1 with illness

Quantity of medical treatment	Your fee-for-service payment (in Taler)	Your bonus payment (in Taler)	Your costs (in Taler)	Your profit (in Taler)	Benefit of the patient with illness and severity (in Taler)

Which quantity of medical treatment do you provide?

Your decision:

OK

[CAP+P4P:]

Patient 1 with illness

Quantity of medical treatment	Your capitation payment (in Taler)	Your bonus payment (in Taler)	Your costs (in Taler)	Your profit (in Taler)	Benefit of the patient with illness and severity (in Taler)

Which quantity of medical treatment do you want to provide?

Your decision:

OK

Comprehension Questions

Prior to the decision rounds, we kindly ask you to answer a few comprehension questions. They are intended to help you familiarize yourself with the decision situations. If you have any questions about this, please raise your hand. Part *II* of the experiment will begin once all participants have answered all comprehension questions correctly.

Finally, we kindly ask you to not talk to anyone about the content of this session in order to prevent influencing other participants after you. Thank you for your cooperation!

B.1.5 Comprehension questions

Experimental condition 1: FFS and FFS+P4P

FFS: (For each of the situations 1. to 4. below, please answer the following questions.)

- What is the fee-for-service payment?
- What are the costs?
- What is the profit?
- What is the patient benefit?

Quantity of medical treatment	Your fee-for-service payment (in Taler)	Your costs (in Taler)	Your profit (in Taler)	Benefit of the patient with illness F and severity y (in Taler)
0	0.00	0.00	0.00	15.00
1	4.00	0.20	3.80	16.00
2	8.00	0.80	7.20	17.00
3	12.00	1.80	10.20	18.00
4	16.00	3.20	12.80	19.00
5	20.00	5.00	15.00	20.00
6	24.00	7.20	16.80	19.00
7	28.00	9.80	18.20	18.00
8	32.00	12.80	19.20	17.00
9	36.00	16.20	19.80	16.00
10	40.00	20.00	20.00	15.00

1. Assume that a physician wants to provide 2 quantities of medical treatment for the patient depicted above.
2. Assume that a physician wants to provide 9 quantities of medical treatment for the patient depicted above.

Quantity of medical treatment	Your fee-for-service payment (in Taler)	Your costs (in Taler)	Your profit (in Taler)	Benefit of the patient with illness G and severity z (in Taler)
0	0.00	0.00	0.00	10.00
1	4.00	0.20	3.80	12.00
2	8.00	0.80	7.20	14.00
3	12.00	1.80	10.20	16.00
4	16.00	3.20	12.80	18.00
5	20.00	5.00	15.00	20.00
6	24.00	7.20	16.80	22.00
7	28.00	9.80	18.20	24.00
8	32.00	12.80	19.20	22.00
9	36.00	16.20	19.80	20.00
10	40.00	20.00	20.00	18.00

3. Assume that a physician wants to provide 2 quantities of medical treatment for the patient depicted above.
4. Assume that a physician wants to provide 9 quantities of medical treatment for the patient depicted above.

FFS+P4P: (For each of the situations 1. to 4. below, please answer the following questions.)

- What is the fee-for-service payment?
- What is the bonus payment?
- What are the costs?
- What is the profit?
- What is the patient benefit?

Quantity of medical treatment	Your fee-for-service payment (in Taler)	Your bonus payment (in Taler)	Your costs (in Taler)	Your profit (in Taler)	Benefit of the patient with illness F and severity y (in Taler)
0	0.00	0.00	0.00	0.00	15.00
1	4.00	0.00	0.20	3.80	16.00
2	8.00	0.00	0.80	7.20	17.00
3	12.00	0.00	1.80	10.20	18.00
4	16.00	7.20	3.20	20.00	19.00
5	20.00	7.20	5.00	22.20	20.00
6	24.00	7.20	7.20	24.00	19.00
7	28.00	0.00	9.80	18.20	18.00
8	32.00	0.00	12.80	19.20	17.00
9	36.00	0.00	16.20	19.80	16.00
10	40.00	0.00	20.00	20.00	15.00

1. Assume that a physician wants to provide 1 quantity of medical treatment for the patient depicted above.
2. Assume that a physician wants to provide 8 quantities of medical treatment for the patient depicted above.

Quantity of medical treatment	Your fee-for-service payment (in Taler)	Your bonus payment (in Taler)	Your costs (in Taler)	Your profit (in Taler)	Benefit of the patient with illness G and severity z (in Taler)
0	0.00	0.00	0.00	0.00	10.00
1	4.00	0.00	0.20	3.80	12.00
2	8.00	0.00	0.80	7.20	14.00
3	12.00	0.00	1.80	10.20	16.00
4	16.00	0.00	3.20	12.80	18.00
5	20.00	0.00	5.00	15.00	20.00
6	24.00	4.80	7.20	21.60	22.00
7	28.00	4.80	9.80	23.00	24.00
8	32.00	4.80	12.80	24.00	22.00
9	36.00	0.00	16.20	19.80	20.00
10	40.00	0.00	20.00	20.00	18.00

3. Assume that a physician wants to provide 1 quantity of medical treatment for the

patient depicted above.

4. Assume that a physician wants to provide 8 quantities of medical treatment for the patient depicted above.

Experimental condition 2: CAP and CAP+P4P

CAP: (*For each of the situations 1. to 4. below, please answer the following questions.*)

- What is the capitation payment?
- What are the costs?
- What is the profit?
- What is the patient benefit?

Quantity of medical treatment	Your capitation payment (in Taler)	Your costs (in Taler)	Your profit (in Taler)	Benefit of the patient with illness F and severity y (in Taler)
0	20.00	0.00	20.00	15.00
1	20.00	0.20	19.80	16.00
2	20.00	0.80	19.20	17.00
3	20.00	1.80	18.20	18.00
4	20.00	3.20	16.80	19.00
5	20.00	5.00	15.00	20.00
6	20.00	7.20	12.80	19.00
7	20.00	9.80	10.20	18.00
8	20.00	12.80	7.20	17.00
9	20.00	16.20	3.80	16.00
10	20.00	20.00	0.00	15.00

1. Assume that a physician wants to provide 2 quantities of medical treatment for the patient depicted above.
2. Assume that a physician wants to provide 9 quantities of medical treatment for the patient depicted above.

Quantity of medical treatment	Your capitation payment (in Taler)	Your costs (in Taler)	Your profit (in Taler)	Benefit of the patient with illness G and severity z (in Taler)
0	20.00	0.00	20.00	10.00
1	20.00	0.20	19.80	12.00
2	20.00	0.80	19.20	14.00
3	20.00	1.80	18.20	16.00
4	20.00	3.20	16.80	18.00
5	20.00	5.00	15.00	20.00
6	20.00	7.20	12.80	22.00
7	20.00	9.80	10.20	24.00
8	20.00	12.80	7.20	22.00
9	20.00	16.20	3.80	20.00
10	20.00	20.00	0.00	18.00

3. Assume that a physician wants to provide 2 quantities of medical treatment for the patient depicted above.
4. Assume that a physician wants to provide 9 quantities of medical treatment for the patient depicted above.

CAP+P4P: (For each of the situations 1. to 4. below, please answer the following questions.)

- What is the capitation payment?
- What is the bonus payment?
- What are the costs?
- What is the profit?
- What is the patient benefit?

Quantity of medical treatment	Your capitation payment (in Taler)	Your bonus payment (in Taler)	Your costs (in Taler)	Your profit (in Taler)	Benefit of the patient with illness F and severity y (in Taler)
0	20.00	0.00	0.00	20.00	15.00
1	20.00	0.00	0.20	19.80	16.00
2	20.00	0.00	0.80	19.20	17.00
3	20.00	0.00	1.80	18.20	18.00
4	20.00	7.20	3.20	24.00	19.00
5	20.00	7.20	5.00	22.20	20.00
6	20.00	7.20	7.20	20.00	19.00
7	20.00	0.00	9.80	10.20	18.00
8	20.00	0.00	12.80	7.20	17.00
9	20.00	0.00	16.20	3.80	16.00
10	20.00	0.00	20.00	0.00	15.00

1. Assume that a physician wants to provide 1 quantity of medical treatment for the patient depicted above.
2. Assume that a physician wants to provide 8 quantities of medical treatment for the patient depicted above.

Quantity of medical treatment	Your capitation payment (in Taler)	Your bonus payment (in Taler)	Your costs (in Taler)	Your profit (in Taler)	Benefit of the patient with illness G and severity z (in Taler)
0	20.00	0.00	0.00	20.00	10.00
1	20.00	0.00	0.20	19.80	12.00
2	20.00	0.00	0.80	19.20	14.00
3	20.00	0.00	1.80	18.20	16.00
4	20.00	0.00	3.20	16.80	18.00
5	20.00	0.00	5.00	15.00	20.00
6	20.00	11.20	7.20	24.00	22.00
7	20.00	11.20	9.80	21.40	24.00
8	20.00	11.20	12.80	18.40	22.00
9	20.00	0.00	16.20	3.80	20.00
10	20.00	0.00	20.00	0.00	18.00

3. Assume that a physician wants to provide 1 quantity of medical treatment for the

patient depicted above.

4. Assume that a physician wants to provide 8 quantities of medical treatment for the patient depicted above.

B.2 Behavioral predictions

Let physician i choose the quantity of medical services q in order to maximize her utility

$$U_i(q) = \alpha_i H(q) + (1 - \alpha_i) \pi(q), \quad (4.3)$$

with $\alpha_i \in [0, 1)$. α_i is a measure for physician i 's altruism. For a purely profit-maximizing physician, for example, $\alpha_i = 0$. A profit-maximizing physician therefore obtains the highest utility, in the absence of P4P in our experiment, when choosing 10 medical services in FFS and when choosing 0 medical services in CAP.

First, we consider physician i 's behavior under the baseline payment systems, i.e., FFS and CAP. For profits and patient benefits given in our experiment, and the given altruism of physician i , we state the following lemma:⁶⁹

Lemma 1 *Physician i overprovides medical services ($q > q^*$) if $p > q^*/5 + (\alpha_i/(1 - \alpha_i)) \theta$, and she underprovides medical services ($q < q^*$) if $p < q^*/5 - (\alpha_i/(1 - \alpha_i)) \theta$. Otherwise, physician i chooses the patient optimal quantity ($q = q^*$).*

Physician i 's objective function $U_i(q) = \alpha_i H(q) + (1 - \alpha_i) \pi(q)$ is concave. Payment $R(q) = L + pq$ is linear and $-c(q)$ is concave as $c(q)$ is convex, thus $\pi(q)$ is a concave function. As $H(q)$ is also a concave function (as defined in Equation 2.2) and $\alpha_i \geq 0$, it follows that $U_i(q)$ is concave.

Note that as $H(q)$ is not differentiable at $q = q^*$, with $q^* \in (0, 10)$. For $q < q^*$, the first-order condition $U'_i(q) = (1 - \alpha_i) [p - \frac{q}{5}] + \alpha_i \theta$. For $q > q^*$, the first-order condition $U'_i(q) = (1 - \alpha_i) [p - \frac{q}{5}] - \alpha_i \theta$. For $q > q^*$, consider $\lim_{q \rightarrow q^*} U'_i(q) = (1 - \alpha_i) [p - \frac{q^*}{5}] - \alpha_i \theta$. If $p < q^*/5 - (\alpha_i/(1 - \alpha_i)) \theta$, $\lim_{q \rightarrow q^*} U'_i(q)$ is positive. Also, because $U_i(q)$ is concave, $U'_i(q) > 0 \forall q < q^*$. Therefore any q such that $q \leq q^*$ cannot be optimal, i.e., physician i chooses $q > q^*$.

Analogously for $q < q^*$, consider $\lim_{q \rightarrow q^*} U'_i(q) = (1 - \alpha_i) [p - \frac{q^*}{5}] + \alpha_i \theta$. If $p < q^*/5 + (\alpha_i/(1 - \alpha_i)) \theta$, $\lim_{q \rightarrow q^*} U'_i(q)$ is negative. Also because $U_i(q)$ is concave, $U'_i(q) < 0 \forall q > q^*$. Therefore any q such that $q \geq q^*$ cannot be optimal, i.e., physician i chooses $q < q^*$.

⁶⁹Notice that Lemma 1 is a special case of Proposition 1 in Brosig-Koch et al. (2017a). They consider a physician's behavior under mixed payment systems with a weight on a FFS component and a lump-sum CAP.

It directly follows from Lemma 1 that physician i 's provision behavior depends on the severity of illness (i.e., the patient-optimal quantity q^* varying with severity of illness l), the fee for a medical service p , the marginal patient health benefit θ , and α_i , the physician i 's degree of altruism. Intuitively, the higher physician i 's altruism is towards her patient, the lower the degree of non-optimal service provision is. Based on Lemma 1, we expect that FFS induces overprovision of medical services, which decreases in the severity of a patient's illness and in patients' marginal health benefit. On the contrary, we expect that CAP induces underprovision of medical services, which increases in the severity of a patient's illness, it decreases in patients' marginal health benefit.

We now focus on the effect of introducing P4P on physicians' health care service provision. Comparing physician i 's provision behavior between FFS (CAP) and FFS+P4P (CAP+P4P), we state the following proposition:

Proposition 1 *Performance pay linked to the optimal patient's health benefit reduces physicians' overprovision of medical services in fee-for-service and underprovision in capitation.*

Let $q^{\text{Opt.}}$ be a physician's utility-maximizing choice for a patient j under FFS or CAP. Depending on a physician's quantity choice, we distinguish three cases. First, we consider $q^{\text{Opt.}} \in [q^* - 1, q^* + 1]$. As $b_l > 0$, it follows that a physician with $\alpha_i \in [0, 1)$ does not change her behavior since $b_l > 0$ is a constant. Second, we consider $q^{\text{Opt.}} > q^* + 1$. Here, the physician chooses q according to $\max\{U^{II}(q^{\text{Opt.}}), U(q^* + 1) + b_l\}$. That means a physician either does not change her behavior or chooses $q^* + 1$ when P4P has been introduced. Analogously for $q^{\text{Opt.}} < q^* - 1$, the same logic applies.

Intuitively, whether a physician meets the quality threshold ($|q - q^*| \leq 1$) depends on physician i 's degree of altruism towards the patient, according to Lemma 1, counterbalancing the incentive effects in FFS and CAP. For a physician's given altruism with $\alpha_i \in [0, 1)$, introducing P4P, therefore, reduces non-optimal service provision under FFS and CAP. Since former experimental evidence shows that non-optimal service provision is highest for those patients where the difference between for whom the incentive effects in FFS and CAP and the patient's optimal quantity are the most misaligned, \hat{q} i.e., for mild severe ill patients under FFS and high severe ill patients under CAP, the effect sizes of P4P are also likely

to vary between severity types. Thus, for a physician's given altruism with $\alpha_i \in [0, 1)$, we expect a larger effect of P4P on non-optimal service provision with increasing severity of illness under CAP and decreasing severity under FFS.

B.3 Additional analyses

Table B.3.1: Quantities and qualities of medical service provision by patients' health characteristics and payment system

	FFS		FFS+P4P			CAP		CAP+P4P		
	Mean	s.d.	Mean	s.d.	p-value	Mean	s.d.	Mean	s.d.	p-value
A. Quantity of medical services q										
Aggregate	6.69	2.07	5.59	1.66	<0.001	3.32	2.13	4.40	1.71	<0.001
Mild severity	5.69	2.44	3.70	0.64	<0.001	2.23	1.22	2.55	0.83	0.0095
Intermediate severity	6.69	1.74	5.58	0.53	<0.001	3.35	1.84	4.38	0.72	<0.001
High severity	7.69	1.36	7.50	0.56	0.0759	4.38	2.55	6.27	0.81	<0.001
Low marginal health benefit	6.69	2.12	5.61	1.67	<0.001	3.23	2.18	4.37	1.69	<0.001
High marginal health benefit	6.70	1.96	5.56	1.63	<0.001	3.49	2.03	4.47	1.74	<0.001
B. Absolute deviation from patient-optimal care ρ										
Aggregate	1.82	1.95	0.63	0.55	<0.001	1.77	2.01	0.65	0.75	<0.001
Mild severity	2.73	2.40	0.74	0.59	<0.001	0.90	1.12	0.59	0.73	0.0035
Intermediate severity	1.79	1.64	0.62	0.49	<0.001	1.75	1.75	0.62	0.72	<0.001
High severity	0.95	1.19	0.53	0.54	<0.001	2.66	2.51	0.75	0.78	<0.001
Low marginal health benefit	1.85	2.00	0.64	0.54	<0.001	1.84	2.08	0.67	0.74	<0.001
High marginal health benefit	1.76	1.85	0.60	0.56	<0.001	1.64	1.86	0.63	0.77	<0.001
C. Proportional health benefit \hat{H}										
Aggregate	0.71	0.31	0.90	0.09	<0.001	0.71	0.32	0.90	0.12	<0.001
Mild severity	0.61	0.34	0.89	0.08	<0.001	0.87	0.16	0.92	0.10	0.0039
Intermediate severity	0.64	0.33	0.88	0.1	<0.001	0.65	0.35	0.88	0.14	<0.001
High severity	0.86	0.17	0.92	0.08	<0.001	0.62	0.36	0.89	0.11	<0.001
High marginal health benefit	0.71	0.30	0.90	0.09	<0.001	0.73	0.30	0.90	0.13	<0.001
Observations	468		468			495		495		
Subjects	52		52			55		55		

Notes. This table shows descriptive statistics on the quantity and quality of medical service provision for each payment condition, at the aggregate level and differentiated by patients' characteristics (levels of severity of illness and marginal health benefit). Two-sided p-values are shown for Wilcoxon signid rank tests for differences in the quantity and quality measures across non-blended (FFS or CAP) and blended payment system (FFS+P4P or CAP+P4P, respectively).

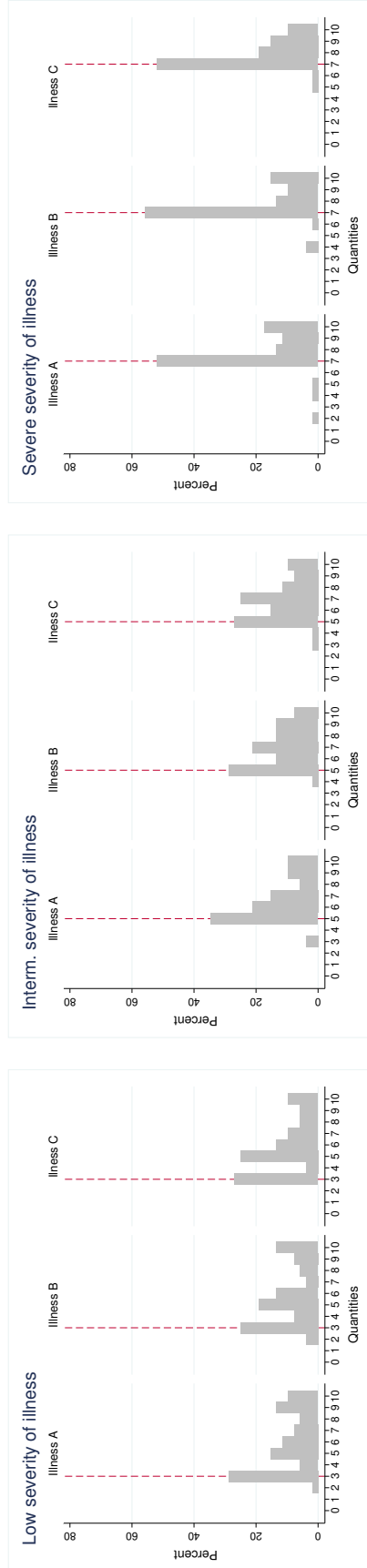
Table B.3.2: Quantity and quality of health care provision by payment system, illness, and severity of illness

	A. Quantity of medical services q			B. Absolute deviation from optimal care ρ			C. Proportional health benefit \hat{H}					
	unblended	+P4P		unblended	+P4P		unblended	+P4P				
		%-change	p -value		%-change	p -value		%-change	p -value			
Fee-For-Service												
Mild severity of illness												
Illness A	5.77 (2.53)	3.65 (0.52)	-0.37	<0.001	2.81 (2.49)	0.69 (0.47)	-0.75	<0.001	0.60 (0.36)	0.90 (0.07)	0.50	<0.001
Illness B	5.67 (2.55)	3.73 (0.69)	-0.34	<0.001	2.75 (2.46)	0.77 (0.65)	-0.72	<0.001	0.61 (0.35)	0.89 (0.09)	0.47	<0.001
Illness C	5.63 (2.28)	3.71 (0.70)	-0.34	<0.001	2.63 (2.28)	0.74 (0.59)	-0.72	<0.001	0.62 (0.33)	0.89 (0.09)	0.43	<0.001
Interm. severity of illness												
Illness A	6.48 (1.82)	5.60 (0.53)	-0.14	<0.001	1.63 (1.68)	0.63 (0.49)	-0.61	<0.001	0.67 (0.34)	0.87 (0.10)	0.30	<0.001
Illness B	6.87 (1.69)	5.58 (0.54)	-0.19	<0.001	1.90 (1.65)	0.62 (0.49)	-0.67	<0.001	0.62 (0.33)	0.88 (0.10)	0.42	<0.001
Illness C	6.73 (1.73)	5.56 (0.54)	-0.17	<0.001	1.85 (1.60)	0.6 (0.5)	-0.68	<0.001	0.63 (0.32)	0.88 (0.10)	0.40	<0.001
Severe severity of illness												
Illness A	7.69 (1.57)	7.58 (0.64)	-0.01	0.5560	1.08 (1.33)	0.62 (0.6)	-0.43	0.0422	0.85 (0.19)	0.91 (0.09)	0.08	0.0422
Illness B	7.65 (1.37)	7.50 (0.50)	-0.02	0.4757	0.92 (1.20)	0.5 (0.50)	-0.46	0.0388	0.87 (0.17)	0.93 (0.07)	0.07	0.0655
Illness C	7.73 (1.12)	7.42 (0.54)	-0.04	0.0691	0.85 (1.04)	0.46 (0.50)	-0.46	0.0205	0.88 (0.15)	0.93 (0.07)	0.06	0.0205
Capitation												
Mild severity of illness												
Illness A	2.07 (1.15)	2.47 (0.69)	0.19	0.0088	0.93 (1.15)	0.6 (0.63)	-0.35	0.0592	0.87 (0.16)	0.91 (0.09)	0.05	0.0592
Illness B	2.20 (1.24)	2.55 (0.72)	0.16	0.1388	0.95 (1.13)	0.56 (0.63)	-0.32	0.0449	0.86 (0.16)	0.92 (0.09)	0.06	0.0717
Illness C	2.42 (1.26)	2.64 (1.04)	0.09	0.6947	0.84 (1.10)	0.62 (0.91)	-0.26	0.2786	0.88 (0.16)	0.91 (0.13)	0.04	0.2786
Interm. severity of illness												
Illness A	3.35 (1.94)	4.36 (0.78)	0.30	<0.001	1.76 (1.84)	0.64 (0.78)	-0.64	<0.001	0.65 (0.37)	0.87 (0.16)	0.35	<0.001
Illness B	3.24 (1.82)	4.40 (0.60)	0.36	<0.001	1.84 (1.74)	0.6 (0.6)	-0.67	<0.001	0.63 (0.35)	0.88 (0.12)	0.39	<0.001
Illness C	3.35 (1.80)	4.38 (0.78)	0.31	<0.001	1.65 (1.70)	0.62 (0.78)	-0.62	<0.001	0.67 (0.34)	0.88 (0.16)	0.31	<0.001
Severe severity of illness												
Illness A	4.27 (2.76)	6.25 (0.78)	0.46	<0.001	2.8 (2.68)	0.78 (0.74)	-0.72	<0.001	0.60 (0.38)	0.89 (0.11)	0.48	<0.001
Illness B	4.25 (2.59)	6.18 (0.98)	0.45	<0.001	2.75 (2.59)	0.82 (0.98)	-0.70	<0.001	0.61 (0.37)	0.88 (0.14)	0.45	<0.001
Illness C	4.60 (2.30)	6.38 (0.62)	0.39	<0.001	2.44 (2.26)	0.65 (0.58)	-0.73	<0.001	0.88 (0.16)	0.91 (0.08)	0.03	<0.001

Notes. This table shows descriptive statistics on the quantities and quality of medical service provision at the level of payment systems, illnesses, severities of illness (means and standard deviations in brackets). 23 non-medical, 22 medical students and 10 physicians decide in the CAP (amounting to a total 990 observations) and 20 non-medical, 22 medical students and 10 physicians in the FFS condition (936 observations). Two-sided p -values are shown for Wilcoxon signed rank tests for matched samples.

Figure B.3.1: Distributions of subjects' quantity choice by severity of illness under different payments scheme

(a) Fee-For-Service



(b) Fee-For-Service + Performance Pay

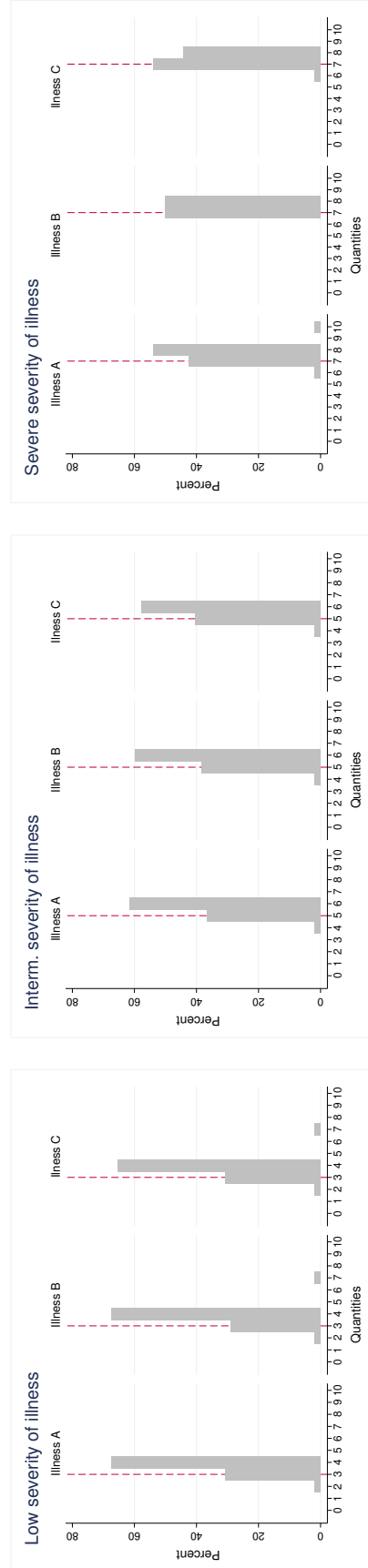
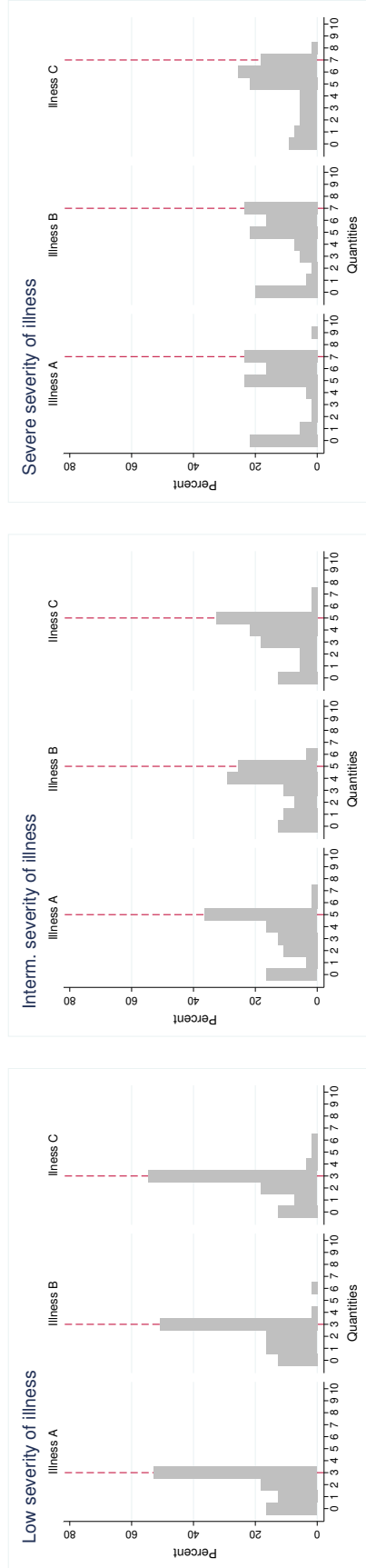


Figure B.3.1: Distributions of subjects' quantity choice by severity of illness under different payments scheme (continued)

(c) Capitation



(d) Capitation + Performance Pay

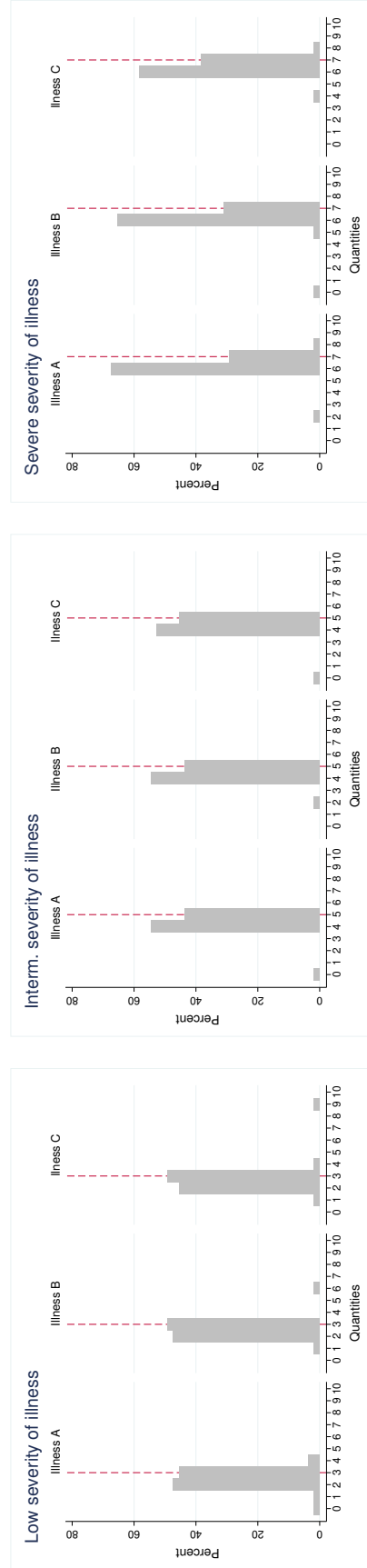


Table B.3.3: Regression models on the effect on quantity and quality between baseline CAP and FFS

	A. Quantity of medical services q		B. Absolute deviation from optimal care ρ		C. Proportional health benefit \hat{H}	
	(1)	(2)	(3)	(4)	(5)	(6)
CAP	-3.375*** (0.320)	-3.443*** (0.309)	-0.053 (0.300)	-0.063 (0.267)	0.009 (0.048)	0.009 (0.042)
INTERMSEV	1.059*** (0.100)	1.059*** (0.101)	-0.019 (0.133)	-0.019 (0.133)	-0.099*** (0.023)	-0.100*** (0.023)
HIGHSEV	2.075*** (0.167)	2.075*** (0.168)	0.037 (0.247)	0.037 (0.248)	-0.006 (0.035)	-0.009 (0.035)
HIGHMHB	0.139** (0.057)	0.139** (0.057)	-0.136** (0.056)	-0.136** (0.056)	0.020** (0.009)	0.021** (0.009)
Medical students		-0.471 (0.358)		-0.365 (0.346)		0.050 (0.057)
Physicians		-1.212** (0.505)		-1.442*** (0.382)		0.236*** (0.057)
Male		-0.468 (0.346)		0.340 (0.317)		-0.057 (0.052)
Extraversion		0.444 (0.383)		-0.077 (0.346)		0.010 (0.053)
Neuroticism		0.119 (0.321)		-0.126 (0.315)		0.018 (0.054)
Openness		-0.097 (0.359)		-0.017 (0.322)		0.001 (0.050)
Conscientiousness		0.930** (0.456)		-0.347 (0.409)		0.053 (0.061)
Agreeableness		0.721 (0.450)		-0.728* (0.392)		0.121* (0.063)
Constant	5.601*** (0.280)	5.876*** (0.380)	1.864*** (0.280)	2.284*** (0.360)		
Observations	963	963	963	963	963	963
Observations	107	107	107	107	107	107
(Pseudo) R^2	0.493	0.534	0.001	0.127	0.009	0.066

Notes. This table shows parameter estimates from OLS regressions (Panel A and B) and average marginal effects from fractional probit response regressions (Panel C). Robust standard errors clustered for subjects are shown in parentheses. CAP is dummy indicating that physicians are remunerated by CAP, and 0 otherwise (remunerated by FFS). INTERMSEV and HIGHSEV are dummy variables for intermediate and high severities of illness. HIGHMHB is a dummy for the marginal health benefit being 1 if $\theta = 2$ (high), and 0 otherwise ($\theta = 1$, low). All models control for individual characteristics which comprise gender, medical experience (non-medical student, medical student, physician), and personality traits. The reference category for medical experience is non-medical students. * $p < 0.10$, ** $p < 0.05$, and, *** $p < 0.01$.

Table B.3.4: Regression models on the effect on quantity and quality under FFS conditions without individual control

	A. Quantity of medical services			B. Abs. deviation from optimal care			C. Proportional health benefit		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
P4P	-1.100*** (0.184)			-1.199*** (0.171)			0.191*** (0.027)		
INTERMSEV	1.439*** (0.086)	1.000*** (0.147)	1.439*** (0.086)	-0.529*** (0.085)	-0.936*** (0.148)	-0.529*** (0.085)	0.004 (0.010)	0.020* (0.012)	0.004 (0.010)
HIGHSEV	2.901*** (0.128)	2.000*** (0.229)	2.901*** (0.128)	-0.997*** (0.141)	-1.782*** (0.255)	-0.997*** (0.141)	0.136*** (0.019)	0.188*** (0.026)	0.136*** (0.019)
HIGHMHB	-0.016 (0.053)	-0.016 (0.053)	0.010 (0.088)	-0.054 (0.051)	-0.054 (0.051)	-0.074 (0.086)	0.008 (0.008)	0.008 (0.008)	0.007 (0.010)
P4P × MILDSEV		-1.994*** (0.276)			-1.994*** (0.276)			0.188*** (0.024)	
P4P × INTERMSEV		-1.115*** (0.191)			-1.179*** (0.182)			0.163*** (0.023)	
P4P × HIGHSEV		-0.192 (0.132)			-0.423*** (0.111)			0.077*** (0.017)	
P4P × LOWMHB			-1.083*** (0.195)			-1.212*** (0.179)			0.172*** (0.025)
P4P × HIGHMHB			-1.135*** (0.176)			-1.173*** (0.171)			0.154*** (0.021)
Constant	5.251*** (0.279)	5.698*** (0.327)	5.243*** (0.284)	2.351*** (0.265)	2.749*** (0.319)	2.358*** (0.268)			
Wald test (p -value)									
P4P × sev.									
H_0 : P4P × MILDSEV = P4P × INTERMSEV			<0.001					0.010	
H_0 : P4P × MILDSEV = P4P × HIGHSEV			<0.001					<0.001	
H_0 : P4P × INTERMSEV = P4P × HIGHSEV			<0.001					<0.001	
P4P × MHB									
H_0 : P4P × LOWMHB = P4P × HIGHMHB			0.554			0.657			0.056
Observations	936	936	936	936	936	936	936	936	936
Subjects	52	52	52	52	52	52	52	52	52
(Pseudo) R^2	0.449	0.484	0.449	0.219	0.262	0.219	0.091	0.098	0.091

Notes. This table shows parameter estimates from OLS regressions (Panel A and B) and average marginal effects from fractional probit response regressions (Panel C). Robust standard errors clustered for subjects are shown in parentheses. P4P is a dummy variable indicating the introduction of P4P. INTERMSEV and HIGHSEV are dummy variables for intermediate and high severities of illness. HIGHMHB is a dummy for the marginal health benefit being 1 if $\theta = 2$ (high), and 0 otherwise ($\theta = 1$, low). * $p < 0.10$, ** $p < 0.05$, and, *** $p < 0.01$.

Table B.3.5: Regression models on the effect on quantity and quality under CAP conditions without individual controls

	A. Quantity of medical services			B. Abs. deviation from optimal care			C. Proportional health benefit		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
P4P	1.085*** (0.188)			-1.117*** (0.179)			0.178*** (0.029)		
INTERMSEV	1.473*** (0.100)	1.115*** (0.138)	1.473*** (0.100)	0.436*** (0.073)	0.848*** (0.137)	0.436*** (0.074)	-0.144*** (0.018)	-0.202*** (0.025)	-0.144*** (0.018)
HIGHSEV	2.933*** (0.151)	2.145*** (0.244)	2.933*** (0.151)	0.958*** (0.133)	1.758*** (0.249)	0.958*** (0.133)	-0.149*** (0.020)	-0.227*** (0.030)	-0.149*** (0.020)
HIGHMHB	0.179*** (0.048)	0.179*** (0.048)	0.261*** (0.069)	-0.115** (0.044)	-0.115** (0.044)	-0.194** (0.073)	0.017** (0.007)	0.017** (0.007)	0.023*** (0.009)
P4P × MILDSEV		0.321** (0.140)			-0.309*** (0.108)			0.057*** (0.018)	
P4P × INTERMSEV		1.036*** (0.200)			-1.133*** (0.188)			0.158*** (0.024)	
P4P × HIGHSEV		1.897*** (0.295)			-1.909*** (0.291)			0.182*** (0.026)	
P4P × LOWMHB			1.139*** (0.195)			-1.170*** (0.188)			0.166*** (0.026)
P4P × HIGHMHB			0.976*** (0.192)			-1.012*** (0.172)			0.137*** (0.021)
Constant	1.789*** (0.180)	2.171*** (0.147)	1.762*** (0.183)	1.345*** (0.173)	0.941*** (0.133)	1.372*** (0.178)			
Wald test (p -value)									
P4P × sev.									
H_0 : P4P × MILDSEV = P4P × INTERMSEV		<0.001			<0.001			<0.001	
H_0 : P4P × MILDSEV = P4P × HIGHSEV		<0.001			<0.001			<0.001	
H_0 : P4P × INTERMSEV = P4P × HIGHSEV		<0.001			<0.001			0.011	
P4P × MHB									
H_0 : P4P × LOWMHB = P4P × HIGHMHB			0.096			0.048			0.004
Observations	990	990	990	990	990	990	990	990	990
Subjects	55	55	55	55	55	55	55	55	55
(Pseudo) R^2	0.432	0.457	0.432	0.180	0.221	0.180	0.082	0.090	0.082

Notes. This table shows parameter estimates from OLS regressions (Panel A and B) and average marginal effects from fractional probit response regressions (Panel C). Robust standard errors clustered for subjects are shown in parentheses. P4P is a dummy variable indicating the introduction of P4P. INTERMSEV and HIGHSEV are dummy variables for intermediate and high severities of illness. HIGHMHB is a dummy for the marginal health benefit being 1 if $\theta = 2$ (high), and 0 otherwise ($\theta = 1$, low). * $p < 0.10$, ** $p < 0.05$, and, *** $p < 0.01$.

Table B.3.6: Regression models on the effect on quantity and quality under FFS conditions with the full list of covariates

Method Model	A. Quantity of medical services q			B. Absolute deviation from optimal care ρ			C. Proportional health benefit \hat{H}		
	OLS (1)	OLS (2)	OLS (3)	OLS (4)	OLS (5)	OLS (6)	Frac. Probit (7)	Frac. Probit (8)	Frac. Probit (9)
P4P	-1.100*** (0.185)			-1.199*** (0.172)			0.189*** (0.025)		
INTERMSEV	1.439*** (0.086)	1.000*** (0.147)	1.439*** (0.086)	-0.529*** (0.086)	-0.936*** (0.149)	-0.529*** (0.086)	0.004 (0.010)	0.019 (0.012)	0.004 (0.010)
HIGHSEV	2.901*** (0.129)	2.000*** (0.230)	2.901*** (0.129)	-0.997*** (0.141)	-1.782*** (0.256)	0.997*** (0.141)	0.133*** (0.016)	0.184*** (0.023)	0.133*** (0.016)
HIGHMHB	-0.016 (0.053)	-0.016 (0.053)	0.010 (0.089)	-0.054 (0.051)	-0.054 (0.051)	-0.074 (0.087)	0.009 (0.008)	0.009 (0.008)	0.008 (0.011)
P4P×MILDSEV		-1.994*** (0.277)			-1.994*** (0.277)			0.187*** (0.021)	
P4P×INTERMSEV		-1.115*** (0.192)			-1.179*** (0.183)			0.162*** (0.021)	
P4P×HIGHSEV		-0.192 (0.132)			-0.423*** (0.111)			0.079*** (0.017)	
P4P×LOWMHB			-1.083*** (0.195)			-1.212*** (0.179)		0.171*** (0.022)	
P4P×HIGHMHB			-1.135*** (0.177)			-1.173*** (0.172)		0.153*** (0.018)	
Medical students									
Physicians	-0.402 (0.306)	-0.402 (0.307)	-0.402 (0.307)	-0.388 (0.299)	-0.388 (0.299)	-0.388 (0.299)	0.064 (0.051)	0.063 (0.051)	0.064 (0.051)
Male	-1.909*** (0.270)	-1.909*** (0.271)	-1.909*** (0.270)	-1.520*** (0.266)	-1.520*** (0.267)	-1.520*** (0.267)	0.230*** (0.041)	0.228*** (0.040)	0.230*** (0.041)
Extraversion	0.227 (0.286)	0.227 (0.286)	0.227 (0.286)	0.275 (0.298)	0.275 (0.299)	0.275 (0.299)	-0.002 (0.035)	-0.002 (0.035)	-0.002 (0.035)
Neuroticism	0.037 (0.255)	0.037 (0.255)	0.037 (0.255)	0.274 (0.274)	0.274 (0.274)	0.274 (0.274)	-0.042 (0.045)	-0.042 (0.045)	-0.042 (0.045)
Openness	-0.151 (0.287)	-0.151 (0.287)	-0.151 (0.287)	-0.021 (0.292)	-0.021 (0.293)	-0.021 (0.292)	-0.046 (0.044)	-0.046 (0.044)	-0.046 (0.044)
Conscientiousness	0.477 (0.316)	0.477 (0.317)	0.477 (0.316)	0.428 (0.325)	0.428 (0.326)	0.428 (0.326)	-0.065 (0.051)	-0.065 (0.051)	-0.065 (0.051)
Agreeableness	0.097 (0.306)	0.097 (0.307)	0.097 (0.306)	0.136 (0.315)	0.136 (0.315)	0.136 (0.315)	-0.030 (0.053)	-0.030 (0.053)	-0.030 (0.053)
Constant	5.623*** (0.315)	6.070*** (0.350)	5.615*** (0.318)	2.621*** (0.315)	3.019*** (0.354)	2.627*** (0.317)			
Wald test (p -value)									
H_0 : P4P×MILDSEV = P4P×INTERMSEV	<0.001			<0.001			0.010		
H_0 : P4P×MILDSEV = P4P×HIGHSEV	<0.001			<0.001			<0.001		
H_0 : P4P×INTERMSEV = P4P×HIGHSEV	<0.001			<0.001			<0.001		
H_0 : P4P×LOWMHB = P4P×HIGHMHB	0.556			0.658			0.062		
Observations	936	936	936	936	936	936	936	936	936
Subjects	52	52	52	52	52	52	52	52	52
(Pseudo) R^2	0.563	0.599	0.563	0.336	0.379	0.336	0.150	0.157	0.150

Notes. This table shows parameter estimates from OLS regressions (Panel A and B) and average marginal effects from fractional probit response regressions (Panel C). Robust standard errors clustered for subjects are shown in parentheses. P4P is a dummy variable indicating the introduction of P4P. INTERMSEV and HIGHSEV are dummy variables for intermediate and high severities of illness. HIGHMHB is a dummy for the marginal health benefit being 1 if $\theta = 2$ (high), and 0 otherwise ($\theta = 1$, low). All models control for individual characteristics which comprise gender, medical experience (non-medical student, medical student, physician), and personality traits. The reference category for medical experience is non-medical students. * $p < 0.10$, ** $p < 0.05$, and, *** $p < 0.01$.

Table B.3.7: Regression models on the effect on quantity and quality under CAP conditions with the full list of covariates

Method Model	A. Quantity of medical services q			B. Absolute deviation from optimal care ρ			C. Proportional health benefit H		
	OLS (1)	OLS (2)	OLS (3)	OLS (4)	OLS (5)	OLS (6)	Frac. Probit (7)	Frac. Probit (8)	Frac. Probit (9)
P4P	1.085*** (0.189)			-1.117*** (0.180)			0.175*** (0.026)		
INTERMSEV	1.473*** (0.100)	1.115*** (0.139)	1.473*** (0.100)	0.436*** (0.074)	0.848*** (0.138)	0.436*** (0.074)	-0.143*** (0.016)	-0.201*** (0.024)	-0.143*** (0.016)
HIGHSEV	2.933*** (0.151)	2.145*** (0.245)	2.933*** (0.151)	0.958*** (0.134)	1.758*** (0.250)	0.958*** (0.134)	-0.149*** (0.019)	-0.227*** (0.028)	-0.149*** (0.019)
HIGHMHB	0.179*** (0.048)	0.179*** (0.048)	0.261*** (0.069)	-0.115*** (0.044)	0.115*** (0.044)	-0.194*** (0.073)	0.017*** (0.007)	0.017*** (0.007)	0.024*** (0.009)
P4P×MILDSEV		0.321** (0.140)			-0.309*** (0.108)			0.055*** (0.017)	
P4P×INTERMSEV		1.036*** (0.201)			-1.133*** (0.189)			0.157*** (0.021)	
P4P×HIGHSEV		1.897*** (0.296)			-1.909*** (0.292)			0.180*** (0.023)	
P4P×LOWMHB			1.139*** (0.195)			-1.170*** (0.188)			0.165*** (0.024)
P4P×HIGHMHB			0.976*** (0.193)			-1.012*** (0.173)			0.135*** (0.018)
Medical students									
Physicians	-0.165 (0.281)	-0.165 (0.281)	-0.165 (0.281)	0.192 (0.280)	0.192 (0.281)	0.192 (0.281)	-0.038 (0.046)	-0.039 (0.046)	-0.038 (0.046)
Male	-0.436* (0.258)	-0.436* (0.258)	-0.436* (0.258)	0.385 (0.263)	0.385 (0.263)	0.385 (0.263)	-0.057 (0.045)	-0.057 (0.045)	-0.057 (0.045)
Extraversion	0.497* (0.283)	0.497* (0.283)	0.497* (0.283)	-0.430 (0.296)	-0.430 (0.296)	-0.430 (0.296)	0.059 (0.045)	0.059 (0.045)	0.059 (0.045)
Neuroticism	0.354 (0.213)	0.354 (0.213)	0.354 (0.213)	-0.301 (0.226)	-0.301 (0.226)	-0.301 (0.226)	0.048 (0.041)	0.047 (0.040)	0.048 (0.041)
Openness	-0.257 (0.237)	-0.257 (0.237)	-0.257 (0.237)	0.198 (0.252)	0.198 (0.252)	0.198 (0.252)	-0.036 (0.039)	-0.035 (0.039)	-0.036 (0.039)
Conscientiousness	0.622** (0.305)	0.622** (0.305)	0.622** (0.305)	-0.690** (0.305)	-0.690** (0.305)	-0.690** (0.305)	0.098** (0.047)	0.097** (0.046)	0.098** (0.047)
Agreeableness	0.874*** (0.300)	0.874*** (0.300)	0.874*** (0.300)	-0.766** (0.297)	-0.766** (0.297)	-0.766** (0.297)	0.121** (0.048)	0.120** (0.048)	0.121** (0.048)
Constant	1.725*** (0.250)	2.107*** (0.231)	1.698*** (0.254)	1.440*** (0.242)	1.036*** (0.227)	1.466*** (0.247)			
Wald test (p -value)									
H_0 : P4P×MILDSEV=P4P×INTERMSEV									
H_0 : P4P×MILDSEV=P4P×HIGHSEV									
H_0 : P4P×INTERMSEV=P4P×HIGHSEV									
H_0 : P4P×LOWMHB=P4P×HIGHMHB			0.097			0.049			0.004
Observations	990	990	990	990	990	990	990	990	990
Subjects	55	55	55	55	55	55	55	55	55
(Pseudo) R^2	0.509	0.534	0.509	0.287	0.328	0.287	0.131	0.140	0.131

Notes. This table shows parameter estimates from OLS regressions (Panel A and B) and average marginal effects from fractional probit response regressions (Panel C). Robust standard errors clustered for subjects are shown in parentheses. P4P is a dummy variable indicating the introduction of P4P. INTERMSEV and HIGHSEV are dummy variables for intermediate and high severities of illness. HIGHMHB is a dummy for the marginal health benefit being 1 if $\theta = 2$ (high), and 0 otherwise ($\theta = 1$, low). All models control for individual characteristics which comprise gender, medical experience (non-medical student, medical student, physician), and personality traits. The reference category for medical experience is non-medical students. * $p < 0.10$, ** $p < 0.05$, and, *** $p < 0.01$.

Table B.3.8: Comparison of effects of blended performance pay systems splitted by marginal health benefit

Method Model	A. Absolute deviation from patient-optimal care ρ		B. Proportional health benefit \bar{H}	
	OLS (1)	OLS (2)	Frac. Probit (3)	Frac. Probit (4)
CAP	-0.013 (0.309)	-0.018 (0.283)	0.001 (0.037)	0.001 (0.034)
INTERMSEV	-0.033 (0.073)	-0.033 (0.073)	-0.063*** (0.013)	-0.063*** (0.013)
HIGHSEV	0.008 (0.135)	0.008 (0.136)	-0.001 (0.019)	-0.003 (0.019)
HIGHMHB	-0.074 (0.086)	-0.074 (0.086)	0.007 (0.010)	0.008 (0.010)
CAP \times HIGHMHB	-0.120 (0.112)	-0.120 (0.113)	0.016 (0.013)	0.017 (0.013)
CAP+P4P \times LowMHB	-1.170*** (0.187)	-1.170*** (0.187)	0.151*** (0.019)	0.150*** (0.018)
FFS+P4P \times LowMHB	-1.212*** (0.178)	-1.212*** (0.178)	0.154*** (0.017)	0.154*** (0.016)
CAP+P4P \times HIGHMHB	-1.012*** (0.172)	-1.012*** (0.172)	0.133*** (0.016)	0.131*** (0.015)
FFS+P4P \times HIGHMHB	-1.173*** (0.170)	-1.173*** (0.171)	0.146*** (0.015)	0.144*** (0.014)
Medical students		-0.136 (0.215)		0.016 (0.035)
Physicians		-0.894*** (0.233)		0.145*** (0.033)
Male		0.185 (0.189)		-0.030 (0.031)
Extraversion		0.008 (0.213)		-0.004 (0.032)
Neuroticism		-0.051 (0.195)		0.006 (0.033)
Openness		0.078 (0.198)		-0.016 (0.030)
Conscientiousness		-0.199 (0.253)		0.029 (0.037)
Agreeableness		-0.470** (0.236)		0.079** (0.037)
Constant	1.858*** (0.245)	2.048*** (0.274)		
Individual controls	No	Yes	No	Yes
Wald tests (p -value):				
H_0 : CAP+P4P \times LowMBH = FFS+P4P \times LowMHB	0.507	0.508	0.425	0.404
H_0 : CAP+P4P \times HIGHMHB = FFS+P4P \times HIGHMHB	0.871	0.872	0.863	0.856
Observations	1926	1926	1926	1926
Subjects	107	107	107	107
(Pseudo) R^2	0.135	0.207	0.064	0.099

Notes. For Panel A, OLS estimates are reported with robust standard errors clustered for subjects (in brackets). For Panel B, average marginal effects based on a fractional probit response model are reported with robust standard errors clustered for subjects (in brackets). CAP = 1 if physicians are remunerated by CAP, and = 0 otherwise (by FFS). P4P is a dummy variable indicating the introduction of P4P. INTERMSEV and HIGHSEV are dummy variables for intermediate and high severities of illness. HIGHMHB is a dummy for the marginal health benefit being 1 if $\theta = 2$ (high), and 0 otherwise ($\theta = 1$, low). Controls for subjects' individual characteristics comprise gender, medical experience (non-medical student, medical student, physician), and personality traits (Big Five Inventory). The reference category for medical experience is non-medical students. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table B.3.9: Comparison of effects of blended performance pay systems with the full list of covariates

Method Model	A. Absolute deviation from patient-optimal care ρ		B. Proportional health benefit \hat{H}	
	OLS (1)	OLS (2)	Frac. Probit (3)	Frac. Probit (4)
CAP	-1.828*** (0.342)	-1.832*** (0.310)	0.210*** (0.037)	0.209*** (0.032)
INTERMSEV	-0.936*** (0.147)	-0.936*** (0.148)	0.020* (0.012)	0.020 (0.012)
HIGHSEV	-1.782*** (0.253)	-1.782*** (0.254)	0.182*** (0.021)	0.178*** (0.020)
HIGHMHB	-0.086** (0.033)	-0.086** (0.034)	0.013** (0.005)	0.013** (0.005)
CAP×INTERMSEV	1.784*** (0.201)	1.784*** (0.201)	-0.243*** (0.029)	-0.241*** (0.028)
CAP×HIGHSEV	3.540*** (0.354)	3.540*** (0.355)	-0.492*** (0.033)	-0.484*** (0.033)
CAP+P4P×MILDSEV	-0.309*** (0.107)	-0.309*** (0.107)	0.056*** (0.017)	0.054*** (0.016)
FFS+P4P×MILDSEV	-1.994*** (0.274)	-1.994*** (0.275)	0.171*** (0.017)	0.170*** (0.016)
CAP+P4P×INTERMSEV	-1.133*** (0.187)	-1.133*** (0.188)	0.148*** (0.018)	0.148*** (0.017)
FFS+P4P×INTERMSEV	-1.179*** (0.181)	-1.179*** (0.182)	0.151*** (0.017)	0.150*** (0.016)
CAP+P4P×HIGHSEV	-1.909*** (0.289)	-1.909*** (0.290)	0.168*** (0.018)	0.167*** (0.017)
FFS+P4P×HIGHSEV	-0.423*** (0.110)	-0.423*** (0.111)	0.075*** (0.015)	0.077*** (0.015)
Medical students		-0.136 (0.215)		0.015 (0.035)
Physicians		-0.894*** (0.233)		0.142*** (0.033)
Male		0.185 (0.189)		-0.031 (0.030)
Extraversion		0.008 (0.214)		-0.004 (0.032)
Neuroticism		-0.051 (0.195)		0.005 (0.033)
Openness		0.078 (0.198)		-0.016 (0.030)
Conscientiousness		-0.199 (0.253)		0.029 (0.037)
Agreeableness		-0.470** (0.236)		0.078** (0.037)
Constant	2.759*** (0.316)	2.950*** (0.324)		
Individual controls	No	Yes	No	Yes
Wald tests (p -value):				
H_0 : CAP+P4P×MILDSEV = FFS+P4P×MILDSEV	<0.001	<0.001	<0.001	<0.001
H_0 : CAP+P4P×INTERMSEV = FFS+P4P×INTERMSEV	0.860	0.860	0.872	0.884
H_0 : CAP+P4P×HIGHSEV = FFS+P4P×HIGHSEV	<0.001	<0.001	<0.001	<0.001
Observations	1926	1926	1926	1926
Subjects	107	107	107	107
(Pseudo) R^2	0.240	0.312	0.094	0.129

Notes. For Panel A, OLS estimates are reported with robust standard errors clustered for subjects (in brackets). For Panel B, average marginal effects based on a fractional probit response model are reported with robust standard errors clustered for subjects (in brackets). CAP = 1 if physicians are remunerated by CAP, and = 0 otherwise (by FFS). P4P is a dummy variable indicating the introduction of P4P. INTERMSEV and HIGHSEV are dummy variables for intermediate and high severities of illness. HIGHMHB is a dummy for the marginal health benefit being 1 if $\theta = 2$ (high), and 0 otherwise ($\theta = 1$, low). Controls for subjects' individual characteristics comprise gender, medical experience (non-medical student, medical student, physician), and personality traits (Big Five Inventory). The reference category for medical experience is non-medical students. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

C Appendix to Chapter 3

C.1 Additional information about the experiment

C.1.1 BIG Five personality inventory (BFI-10)

Table C.1.1: Item description of BFI-10 by Rammstedt et al. (2007)

Trait	Item wording
EXTRAVERSION	“I see myself as someone who is reserved.”(R) “I see myself as someone who is outgoing, sociable.”
NEUROTICISM	“I see myself as someone who is relaxed, handles stress well.” (R) “I see myself as someone who gets nervous easily.”
OPENNESS	“I see myself as someone who has few artistic interests.” (R) “I see myself as someone who has an active imagination.”
CONSCIENTIOUSNESS	“I see myself as someone who tends to be lazy.” (R) “I see myself as someone who does a thorough job.”
AGREEABLENESS	“I see myself as someone who tends to find fault with others.” (R) “I see myself as someone who is generally trusting.”

Notes. Subjects express their agreement to each statement on a five point Likert scale with 1 corresponding to disagree strongly, 2 to disagree a little, 3 to neutral, 4 to agree a little, and 5 to agree strongly. Items marked (R) are reverse coded. Two out of the ten statements form one trait.

C.1.2 Sample

While we investigate the effect of P4P at a within-subject level, we vary the underlying payment system (CAP or FFS) between subjects. Note that we are mainly interested in the effect of P4P and its interaction with personality traits considered separately for each payment system rather than in a across-payment comparison. Hence, at a within-subject design, we perform separate subanalyses for both payment systems instead of analyzing both payments systems in one model. Nonetheless, it might be possible that potential variations in the link between personality and performance across payment systems rests upon different samples with respect to personality traits. Table C.1.2 shows statistics for each personality measure for our subsamples. In general, we observe no significant differences in personality measures between our sample under FFS conditions and our sample under CAP conditions, with the exception of statistical weakly significant difference in conscientiousness based on a Mann-Whitney U (MWU) test. Subjects under CAP appear to have slightly higher scores

Table C.1.2: Personality traits by payment systems

	Mean (s.d.)		Difference	<i>p</i> -value	
	FFS	CAP		MWU	KS
EXTRAVERSION	3.683 (0.863)	3.436 (0.788)	0.247	0.114	0.609
NEUROTICISM	2.798 (0.925)	2.745 (1.018)	0.053	0.732	0.995
OPENNESS	3.548 (0.930)	3.627 (0.909)	-0.079	0.596	0.992
CONSCIENTIOUSNESS	3.452 (0.794)	3.691 (0.825)	-0.239	0.097	0.382
AGREEABLENESS	3.115 (0.669)	3.000 (0.758)	0.115	0.495	1.000
Subjects	52	55			

Notes. For each personality trait we report mean responses to two statements representing the trait. Subjects express their agreement to each statement on a five point Likert scale with 1 corresponding to disagree strongly, 2 to disagree a little, 3 to neutral, 4 to agree a little, and 5 to agree strongly. Variable standard deviations are reported in parentheses. *P*-values based on Mann-Whitney-U (MWU) tests for differences between subsamples and based on two-sample Kolmogorov-Smirnov (KS) tests are reported in the second last and in the last column, respectively.

in conscientiousness. When we, however, compare the entire distributions of both samples, *p*-values of Kolmogorov-Smirnov (KS) tests ($p \geq 0.382$) indicate that there is no statistical evidence of a difference in any personality traits between the distributions. Taken together, the random allocation of subjects into either a CAP or FFS payment system has resulted in a fairly balanced samples with respect to all personality traits.

C.2 Additional analyses

Table C.2.3: Regression models on the interaction effects of performance pay and personality traits under CAP, trait by trait analyse

	Absolute deviation from patient-optimal care					
	(1)	(2)	(3)	(4)	(5)	(6)
P4P	-1.167*** (0.215)	-1.097*** (0.179)	-1.315*** (0.223)	-1.550*** (0.254)	-1.117*** (0.174)	-1.716*** (0.220)
EXTRAVERSION	-0.543 (0.504)	-0.430 (0.296)	-0.430 (0.296)	-0.430 (0.296)	-0.430 (0.296)	-0.743 (0.492)
NEUROTICISM	-0.301 (0.226)	-0.381 (0.348)	-0.301 (0.226)	-0.301 (0.226)	-0.301 (0.226)	-0.553 (0.357)
OPENNESS	0.198 (0.252)	0.198 (0.252)	-0.118 (0.404)	0.198 (0.252)	0.198 (0.252)	0.041 (0.392)
CONSCIENTIOUSNESS	-0.690** (0.305)	-0.690** (0.305)	-0.690** (0.305)	-1.316*** (0.471)	-0.690** (0.305)	-1.309*** (0.461)
AGREEABLENESS	-0.766** (0.297)	-0.766** (0.297)	-0.766** (0.297)	-0.766** (0.297)	-1.191** (0.534)	-1.162** (0.496)
P4P × EXTRAVERSION	0.227 (0.513)					0.627 (0.460)
P4P × NEUROTICISM		0.162 (0.316)				0.506 (0.305)
P4P × OPENNESS			0.632 (0.407)			0.314 (0.342)
P4P × CONSCIENTIOUSNESS				1.252*** (0.434)		1.238*** (0.390)
P4P × AGREEABLENESS					0.849 (0.549)	0.791* (0.439)
INTERSEV	0.436*** (0.074)	0.436*** (0.074)	0.436*** (0.074)	0.436*** (0.074)	0.436*** (0.074)	0.436*** (0.074)
HIGHSEV	0.958*** (0.134)	0.958*** (0.134)	0.958*** (0.134)	0.958*** (0.134)	0.958*** (0.134)	0.958*** (0.134)
HIGHMHB	-0.115** (0.044)	-0.115** (0.044)	-0.115** (0.044)	-0.115** (0.044)	-0.115** (0.044)	-0.115** (0.044)
MEDICAL STUDENT	0.192 (0.281)	0.192 (0.281)	0.192 (0.281)	0.192 (0.281)	0.192 (0.281)	0.192 (0.281)
PHYSICIAN	-0.438 (0.323)	-0.438 (0.323)	-0.438 (0.323)	-0.438 (0.323)	-0.438 (0.323)	-0.438 (0.323)
MALE	0.385 (0.263)	0.385 (0.263)	0.385 (0.263)	0.385 (0.263)	0.385 (0.263)	0.385 (0.263)
CONSTANT	1.465*** (0.259)	1.429*** (0.246)	1.539*** (0.254)	1.656*** (0.273)	1.440*** (0.244)	1.739*** (0.267)
Observations	990	990	990	990	990	990
Subjects	55	55	55	55	55	55
R^2	0.287	0.287	0.294	0.312	0.296	0.328
Adjusted R^2	0.278	0.278	0.285	0.303	0.287	0.317

Notes. This table shows estimates from separate trait by trait OLS regressions on the effects of performance pay (P4P) and personality traits on quality of care under a capitation (CAP) baseline payment system. We perform separate trait by trait regression analyses for robustness. All personality traits are measured on a scale from 1 to +1. The coefficients thus reflect the effect of a one-unit change; for example, a change from the theoretical minimum to the (neutral) midpoint or from the midpoint to the theoretical maximum. All models control for patient's health characteristics which comprise severity of illness and patient's marginal health benefit. Intermediate severity of illness = 1 if $l = y$, and = 0 otherwise. High severity of illness = 1 if $l = z$, and = 0 otherwise. High marginal health benefit (MHB) is a dummy for indicating a high value of marginal health benefits (= 1 if $\theta = 2$, for illness C, and = 0 if $\theta = 1$ for illness A, B). We also include further individual subjects' controls in all models which comprise gender and medical experience, i.e., non-medical students (=reference category), medical students and physicians. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table C.2.4: Regression models on the interaction effects of performance pay and personality traits under FFS, trait by trait analyses

	Absolute deviation from patient-optimal care					
	(1)	(2)	(3)	(4)	(5)	(6)
P4P	-1.127*** (0.182)	-1.220*** (0.173)	-1.272*** (0.210)	-1.302*** (0.197)	-1.224*** (0.174)	-1.290*** (0.292)
EXTRAVERSION	0.381 (0.457)	0.275 (0.299)	0.275 (0.299)	0.275 (0.299)	0.275 (0.299)	0.376 (0.473)
NEUROTICISM	0.274 (0.258)	0.382 (0.414)	0.274 (0.258)	0.274 (0.258)	0.274 (0.258)	0.367 (0.400)
OPENNESS	-0.021 (0.292)	-0.021 (0.292)	-0.155 (0.467)	-0.021 (0.292)	-0.021 (0.292)	-0.074 (0.488)
CONSCIENTIOUSNESS	0.428 (0.326)	0.428 (0.326)	0.428 (0.326)	0.200 (0.485)	0.428 (0.326)	0.231 (0.492)
AGREEABLENESS	0.136 (0.315)	0.136 (0.315)	0.136 (0.315)	0.136 (0.315)	-0.079 (0.558)	-0.067 (0.560)
P4P × EXTRAVERSION	-0.211 (0.390)					-0.202 (0.406)
P4P × NEUROTICISM		-0.215 (0.380)				-0.186 (0.339)
P4P × OPENNESS			0.267 (0.399)			0.105 (0.429)
P4P × CONSCIENTIOUSNESS				0.457 (0.390)		0.394 (0.394)
P4P × AGREEABLENESS					0.431 (0.548)	0.406 (0.552)
INTERSEV	-0.529*** (0.086)	-0.529*** (0.086)	-0.529*** (0.086)	-0.529*** (0.086)	-0.529*** (0.086)	-0.529*** (0.086)
HIGHSEV	-0.997*** (0.141)	-0.997*** (0.141)	-0.997*** (0.141)	-0.997*** (0.141)	-0.997*** (0.141)	-0.997*** (0.142)
HIGHMHB	-0.054 (0.051)	-0.054 (0.051)	-0.054 (0.051)	-0.054 (0.051)	-0.054 (0.051)	-0.054 (0.051)
MEDICAL STUDENT	-0.388 (0.299)	-0.388 (0.299)	-0.388 (0.299)	-0.388 (0.299)	-0.388 (0.299)	-0.388 (0.300)
PHYSICIAN	-1.520*** (0.267)	-1.520*** (0.267)	-1.520*** (0.267)	-1.520*** (0.267)	-1.520*** (0.267)	-1.520*** (0.267)
MALE	0.050 (0.222)	0.050 (0.222)	0.050 (0.222)	0.050 (0.222)	0.050 (0.222)	0.050 (0.223)
CONSTANT	2.585*** (0.318)	2.632*** (0.320)	2.658*** (0.340)	2.673*** (0.331)	2.634*** (0.318)	2.667*** (0.376)
Observations	936	936	936	936	936	936
Subjects	52	52	52	52	52	52
R^2	0.330	0.330	0.330	0.332	0.331	0.336
Adjusted R^2	0.320	0.320	0.321	0.323	0.322	0.324

Notes. This table shows estimates from separate trait by trait OLS regressions on the effects of performance pay (P4P) and personality traits on quality of care under a fee-for-service (FFS) baseline payment system. We perform separate trait by trait regression analyses for robustness. All personality traits are measured on a scale from 1 to +1. The coefficients thus reflect the effect of a one-unit change; for example, a change from the theoretical minimum to the (neutral) midpoint or from the midpoint to the theoretical maximum. All models control for patient's health characteristics which comprise severity of illness and patient's marginal health benefit. Intermediate severity of illness = 1 if $l = y$, and = 0 otherwise. High severity of illness = 1 if $l = z$, and = 0 otherwise. High marginal health benefit (MHB) is a dummy for indicating a high value of marginal health benefits (= 1 if $\theta = 2$, for illness C , and = 0 if $\theta = 1$ for illness A, B). We also include further individual subjects' controls in all models which comprise gender and medical experience, i.e., non-medical students (=reference category), medical students and physicians. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table C.2.5: Regression models on several versions of our base model

	A. CAP					B. FFS				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
P4P	-1.117*** (0.180)	-1.716*** (0.220)	-1.716*** (0.220)	-1.716*** (0.220)	-1.716*** (0.220)	-1.199*** (0.172)	-1.290*** (0.291)	-1.290*** (0.291)	-1.290*** (0.291)	-1.290*** (0.292)
INTERSEV	0.436*** (0.074)		0.436*** (0.074)		0.436*** (0.074)	-0.529*** (0.086)		-0.529*** (0.086)		-0.529*** (0.086)
HIGHSEV	0.958*** (0.134)		0.958*** (0.134)		0.958*** (0.134)	-0.997*** (0.141)		-0.997*** (0.141)		-0.997*** (0.142)
HIGHMHB	-0.115** (0.044)		-0.115** (0.044)		-0.115** (0.044)	-0.054 (0.051)		-0.054 (0.051)		-0.054 (0.051)
MEDICAL STUDENTS	0.192 (0.280)			0.192 (0.281)	0.192 (0.281)	-0.388 (0.299)			-0.388 (0.299)	-0.388 (0.300)
PHYSICIAN	-0.438 (0.322)			-0.438 (0.323)	-0.438 (0.323)	-1.520*** (0.266)			-1.520*** (0.267)	-1.520*** (0.267)
MALE	0.385 (0.263)			0.385 (0.263)	0.385 (0.263)	0.050 (0.222)			0.050 (0.222)	0.050 (0.223)
EXTRAVERSION	-0.430 (0.296)	-0.649 (0.522)	-0.649 (0.523)	-0.743 (0.492)	-0.743 (0.492)	0.275 (0.298)	0.441 (0.493)	0.441 (0.494)	0.376 (0.472)	0.376 (0.473)
NEUROTICISM	-0.301 (0.226)	-0.617 (0.374)	-0.617 (0.374)	-0.553 (0.356)	-0.553 (0.357)	0.274 (0.258)	0.313 (0.422)	0.313 (0.423)	0.367 (0.400)	0.367 (0.400)
OPENNESS	0.198 (0.252)	-0.009 (0.425)	-0.009 (0.425)	0.041 (0.391)	0.041 (0.392)	-0.021 (0.292)	0.022 (0.504)	0.022 (0.504)	-0.074 (0.487)	-0.074 (0.488)
CONSCIENTIOUSNESS	-0.690** (0.305)	-1.573*** (0.446)	-1.573*** (0.447)	-1.309*** (0.460)	-1.309*** (0.461)	0.428 (0.325)	-0.438 (0.502)	-0.438 (0.503)	0.231 (0.491)	0.231 (0.492)
AGREEABLENESS	-0.766** (0.297)	-1.099** (0.539)	-1.099** (0.540)	-1.162** (0.496)	-1.162** (0.496)	0.136 (0.315)	-0.402 (0.668)	-0.402 (0.669)	-0.067 (0.559)	-0.067 (0.560)
P4P × EXTRAVERSION		0.627 (0.459)	0.627 (0.460)	0.627 (0.460)	0.627 (0.460)		-0.202 (0.404)	-0.202 (0.405)	-0.202 (0.405)	-0.202 (0.406)
P4P × NEUROTICISM		0.506 (0.304)	0.506 (0.304)	0.506 (0.304)	0.506 (0.305)		-0.186 (0.338)	-0.186 (0.338)	-0.186 (0.338)	-0.186 (0.339)
P4P × OPENNESS		0.314 (0.341)	0.314 (0.341)	0.314 (0.341)	0.314 (0.342)		0.105 (0.428)	0.105 (0.428)	0.105 (0.428)	0.105 (0.429)
P4P × CONSCIENTIOUSNESS		1.238*** (0.389)	1.238*** (0.389)	1.238*** (0.389)	1.238*** (0.390)		0.394 (0.393)	0.394 (0.393)	0.394 (0.393)	0.394 (0.394)
P4P × AGREEABLENESS		0.791* (0.437)	0.791* (0.438)	0.791* (0.438)	0.791* (0.439)		0.406 (0.550)	0.406 (0.551)	0.406 (0.551)	0.406 (0.552)
CONSTANT	1.440*** (0.242)	2.381*** (0.251)	1.955*** (0.220)	2.165*** (0.286)	1.739*** (0.267)	2.621*** (0.315)	1.822*** (0.365)	2.348*** (0.401)	2.140*** (0.349)	2.667*** (0.376)
Observations	990	990	990	990	990	936	936	936	936	936
Subjects	55	55	55	55	55	52	52	52	52	52
R ²	0.287	0.248	0.308	0.269	0.328	0.329	0.173	0.242	0.267	0.336

Notes. This table shows estimates from OLS regressions on the effects of performance pay (P4P) and personality traits on quality of care under capitation (CAP, Panel A) or fee-for-service (FFS, Panel B) baseline payment systems. Robust standard errors clustered for subjects are shown in parentheses. The reported regression estimates stem from several versions of our base model, see Equation (3.1). In Column (1) and (6), we study the average effect of personality traits alone. All remaining models include the interaction terms between personality traits and P4P but vary with respect to the inclusion of control variables. Columns (2) and (7) shows the effects on the quality of care without any subject or patient level controls. In Columns (3) and (7), the model is extended by patients' controls which comprise a patient's severity of illness and patient's marginal health benefit. Intermediate severity of illness = 1 if $l = y$, and = 0 otherwise. High severity of illness = 1 if $l = z$, and = 0 otherwise. High marginal health benefit (MHB) is a dummy for indicating a high value of marginal health benefits (= 1 if $\theta = 2$, for illness C , and = 0 if $\theta = 1$ for illness A, B). In Columns (4) and (8), individual subjects' controls which comprise gender and medical experience, i.e., non-medical students (=reference category), medical students and physicians. Columns (5) and (8) show regression estimates with the full list of covariates for our main model (reported in Table 3.1). All personality traits are measured on a scale from 1 to +1. The coefficients thus reflect the effect of a one-unit change; for example, a change from the theoretical minimum to the (neutral) midpoint or from the midpoint to the theoretical maximum. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table C.2.6: Comparisons of regression models for different regression methods

Method Model	A. CAP			B. FFS		
	OLS (1)	Tobit (2)	REML (3)	OLS (4)	Tobit (5)	REML (6)
P4P	-1.716*** (0.220)	-2.030*** (0.232)	-1.716*** (0.247)	-1.290*** (0.292)	-1.706*** (0.290)	-1.290*** (0.286)
INTERSEV	0.436*** (0.074)	0.706*** (0.125)	0.436*** (0.067)	-0.529*** (0.086)	-0.666*** (0.116)	-0.529*** (0.071)
HIGHSEV	0.958*** (0.134)	1.454*** (0.191)	0.958*** (0.067)	-0.997*** (0.142)	-1.474*** (0.204)	-0.997*** (0.071)
HIGHMHB	-0.115** (0.044)	-0.169* (0.093)	-0.115** (0.058)	-0.054 (0.051)	-0.095 (0.095)	-0.054 (0.061)
MEDICAL STUDENTS	0.192 (0.281)	0.400 (0.451)	0.174 (0.217)	-0.388 (0.300)	-0.498 (0.475)	0.114 (0.127)
PHYSICIAN	-0.438 (0.323)	-0.867 (0.617)	-0.093 (0.282)	-1.520*** (0.267)	-3.130*** (0.623)	-0.441*** (0.156)
MALE	0.385 (0.263)	0.543 (0.459)	-0.105 (0.208)	0.050 (0.223)	0.035 (0.370)	0.061 (0.109)
EXTRAVERSION	-0.743 (0.492)	-0.975 (0.640)	-0.707 (0.568)	0.376 (0.473)	0.416 (0.655)	0.415 (0.523)
P4P×EXTRAVERSION	0.627 (0.460)	0.560 (0.520)	0.627 (0.456)	-0.202 (0.406)	0.101 (0.446)	-0.202 (0.451)
NEUROTICISM	-0.553 (0.357)	-0.658 (0.534)	-0.677 (0.450)	0.367 (0.400)	0.662 (0.573)	0.402 (0.485)
P4P×NEUROTICISM	0.506 (0.305)	0.297 (0.389)	0.506 (0.358)	-0.186 (0.339)	-0.268 (0.401)	-0.186 (0.418)
OPENNESS	0.041 (0.392)	0.048 (0.517)	-0.017 (0.464)	-0.074 (0.488)	-0.106 (0.671)	-0.063 (0.471)
P4P×OPENNESS	0.314 (0.342)	0.720* (0.369)	0.314 (0.373)	0.105 (0.429)	0.279 (0.493)	0.105 (0.406)
CONSCIENTIOUSNESS	-1.309*** (0.461)	-1.535** (0.642)	-1.589*** (0.540)	0.231 (0.492)	0.579 (0.696)	-0.223 (0.547)
P4P×CONSCIENTIOUSNESS	1.238*** (0.390)	1.036** (0.447)	1.238*** (0.424)	0.394 (0.394)	0.378 (0.430)	0.394 (0.469)
AGREEABLENESS	-1.162** (0.496)	-1.620** (0.705)	-1.166** (0.561)	-0.067 (0.560)	0.026 (0.774)	-0.389 (0.635)
P4P×AGREEABLENESS	0.791* (0.439)	0.574 (0.508)	0.791* (0.450)	0.406 (0.552)	0.463 (0.591)	0.406 (0.547)
CONSTANT	1.739*** (0.267)	1.080** (0.454)	1.953*** (0.339)	2.667*** (0.376)	2.706*** (0.511)	2.350*** (0.346)
Observations	990	990	990	936	936	936
Subjects	55	55	55	52	52	52
(Pseudo) R^2	0.328	0.102		0.336	0.124	

Notes. This table shows estimates from regressions on the effects of performance pay (P4P) and personality traits on quality of care under capitation (CAP, Panel A) or fee-for-service (FFS, Panel B) baseline payment systems. For each payment system, we report estimates based on OLS regressions, tobit regressions, and multilevel mixed-effects REML regressions. All models include patients' and subjects' level controls. Patients' controls which comprise a patient's severity of illness and patient's marginal health benefit. Intermediate severity of illness = 1 if $l = y$, and = 0 otherwise. High severity of illness = 1 if $l = z$, and = 0 otherwise. High marginal health benefit (MHB) is a dummy for indicating a high value of marginal health benefits (= 1 if $\theta = 2$, for illness C , and = 0 if $\theta = 1$ for illness A, B). Individual subjects' controls comprise gender and medical experience, i.e., non-medical students (=reference category), medical students and physicians. All personality traits are measured on a scale from 1 to +1. The coefficients thus reflect the effect of a one-unit change; for example, a change from the theoretical minimum to the (neutral) midpoint or from the midpoint to the theoretical maximum. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

D Appendix to Chapter 4

D.1 Additional information about the experiment

D.1.1 Instructions of the experiment

Translated from German, lab version [online version]

Experimental study of the Faculty of Medicine and the Faculty of Management, Economics and Social Sciences at the University of Cologne

Thank you for your willingness to participate in our research study in cooperation between the Faculty of Medicine and Faculty of Management, Economics and Social Sciences at the University of Cologne.

In particular, you are participating in a study on decision-making behavior in the context of experimental economic research. The aim of the study is to gain insights into the behavioral mechanisms and determinants of individual decisions. Overall, the study—consisting of a decision-making experiment and a subsequent questionnaire—will take about 45 minutes of your time.

During the experiment, you and the other participants will be asked to make decisions. You can earn money in the course of the experiment. The amount of your earnings depends on your decisions.

You will also be asked to answer a series of questions. As compensation for answering the questionnaire, you will receive an additional fixed amount of €5.

After having completed the questionnaire, all your earnings from the experiment and the questionnaire will be paid to you in cash. [After completion, all your earnings from the experiment and from the questionnaire will be paid to you in cash.]

With your decisions in the experiment, you determine the amount of your own earnings as well as the benefit for a patient. There are no patients present during the experiment, but the patient benefit resulting from your decisions (expressed in Euro) will be transferred to

the charitable organization *Christoffel Blindenmission Deutschland e.V.*, 64625 Bensheim. They spend the amount exclusively on the treatment of patients with cataracts (an eye disease).

Your participation is voluntary. During the experiment and while filling in the questionnaire, you can interrupt or withdraw from participation at any time without giving reasons and without facing any negative consequences for your person.

During the course of your medical studies, we will ask you again to participate in our study. The success of our research project depends on your regular participation. We would like to include about 80% of your semester students in this study.

The data generated from the experiment and from the questionnaire are analyzed pseudonymously. The data cannot be matched to any specific person. Third parties have no access to the data. The pseudonymous data will be used for scientific research and presentations. This work will be published. Study results are not used commercially. Before starting the experiment, we ask you to generate an individual code on your computer. No conclusions about your person can be drawn from this code. This also applies in the case of publication of the study results. The code serves only as a pseudonym, through which your individual data from several participations can be compiled. The code also serves for the anonymous payment of the earnings from the experiment. The 6-digit code consists of the first letter of your place of birth, the second letter of your mother's first name, the second letter of your own first name, the first letter of your father's first name, and your day of birth (in two digits).

A data custodian is responsible for storing and managing the data generated from the experiment. The data custodian is independent of the research project and is the only person who has access to the entire data set, which—in addition to the data collected in the experiment and questionnaire—contains your personal code and student number. The data custodian does not perform any data analyses and ensures that the people involved in the project receive only the data and the pseudonyms, which are collected in the experiment. This ensures that no conclusions can be drawn from the data about individual persons. The

data set containing the participants' pseudonyms and student numbers will be destroyed after completion of the project (after 6 years). You have the option (without restriction) to revoke your consent to data processing. There is the possibility of viewing the original data through the Ethics Commission. You have the right to information and rectification regarding the data stored about you. It is possible to delete personal data.

Contact:

Prof. Dr. Daniel Wiesen

Department of Business Administration and Health Care Management, University of Cologne
Universitätsstraße 91, 50931 Cologne

Tel. + 49 (0) 221 470 89 171

wiesen@wiso.uni-koeln.de

The study was approved by the Ethics Commission of the Faculty of Medicine of the University of Cologne.

[Declaration of consent] *(In the lab, the consent form was handed out and signed by participants.)*

I have read and understood the information about participation in the experimental study, including the information concerning data security. I know that my participation is voluntary and that I can revoke my consent to participate at any time without suffering any negative consequences for my person.

- ☐ I declare my consent for participation and want to participate
- ☐ I do not want to participate

Generation of the 6-digit code:

To link your answers from different surveys, we would like to ask you to enter the five pieces of information indicated below, in order to generate your personal code. No conclusions about your person can be drawn from this code. It serves solely as a pseudonym, through

which your answers to different surveys can be linked. Furthermore, the code allows the anonymous payment of the earnings from the experiment.

(Notes. If you have umlauts in your name, please enter them as such (e.g., ä not ae). Please use lower case letters.)

- o First letter of your place of birth
- o Second letter of your mother's first name
- o Second letter of your own first name
- o First letter of your father's first name

Answer: _____

(Notes. Please do not enter more letters than the maximum of 4. Any letters exceeding this maximum will be deleted automatically.)

Your day of birth (in two digits)

Answer: _____

Generation of the 6-digit code:

To confirm, please re-enter your personal code:

(Notes. If you have umlauts in your name, please enter them as such (e.g., ä not ae). Please use lower case letters.)

- o First letter of your place of birth
- o Second letter of your mother's first name
- o Second letter of your own first name
- o First letter of your father's first name

Answer: _____

(Notes. Please do not enter more letters than the maximum of 4. Any letters exceeding this maximum will be deleted automatically.)

Your day of birth (in two digits)

Answer: _____

Welcome to our decision-making experiment!

Description of the experiment

You are participating in a study on decision-making behavior in the context of experimental economic research. During the experiment, you and the other participants will be asked to make decisions. You can earn money in the course of the experiment. The amount of your earnings depends on your decisions. Please carefully read through the following description of the experiment. *If you have questions regarding the description or during the experiment, you can signal this by raising your hand at any time. We will come to you and answer your questions personally.*

For the entire duration of the session, it is not allowed to communicate with other participants. If you violate this rule, you will be excluded from the experiment and will not receive any payment.

Please make sure that your mobile phones are switched off and that they are put in the corridor along with your bags and jackets if applicable. From there, you can pick up your belongings when you are finished. [While you are doing the experiment, we would like to ask you not to use any auxiliary devices (internet, cell phone, ...) and not to communicate with the others.]

The experiment comprises 30 rounds.

During the entire experiment you take the role of a physician. In each round, you decide on the treatment of a patient.

For each patient, you choose between two treatment options—Treatment A or Treatment B. With your decision, you determine your earnings in Euros and the benefit for the patient, which is also measured in Euros.

Upon completion of the experiment one of the 30 rounds of the experiment will be randomly

selected. The earnings from that round will be paid to you in cash after the completion of the experiment and the questionnaire.

No There are no patients will be present in this during the experiment (or and no participants in take the role of patients). The patient benefit in Euros, which is determined by from your decision in the randomly selected round, will help an actual patient: the amount of money resulting from the patient benefit you generated with your decision regarding the benefit for a patient will be given donated to the charitable organization Christoffel Blindenmission Deutschland e.V., 64265 Bensheim. The amount is earmarked to facilitate the treatment of patients with cataracts (an eye disease).

☐ I have understood the instructions.

Payment

Information concerning place and time of the payment will be given to you after the completion of the experiment. (*Additionally, an information sheet was handed out to the participants*)

You will receive your payment in an envelope with your personal code on it. To guarantee anonymized payment, you will confirm on a separate receipt with your signature that you have received your payment according to your code. The code is not listed on this separate receipt sheet. Once signed, please put the receipt sheet without the envelope in the provided box.

After the experiment, the proper payment of the amount to the *Christoffel Blindenmission Deutschland e.V.* will be confirmed by a control person. When receiving her payment, the control person will enter the amount of money, which results from the cumulated patient benefit generated in the randomly selected round, in a payment order form. The form will be put in a stamped envelope addressed to the financial administration of the University of Cologne. The control person and another person involved in the study will then together put it into the nearest mailbox. The financial administration of the University of Cologne will then issue the payment of this amount to *Christoffel Blindenmission Deutschland e.V.*.

The role of the control person will be randomly assigned to a participant upon completion of the questionnaire [today]. The control person will receive an additional remuneration of 5€ next to the payment she receives from the experiment and the questionnaire. On a form, the control person will confirm with her signature that she has correctly completed her assigned tasks as described above. A copy of this form as well as a copy of the confirmation of *Christoffel Blindenmission Deutschland e.V.*, indicating the receipt of payment, will be hung up in the information box of the Deanery of Studies of the Faculty of Medicine at Cologne University (Basement, Building 42, Joseph-Stelzmann-Str. 20, 50931 Cologne).

☐ I have read and understood the information regarding payment.

Sample decision situation in the experiment

We would now like to explain the decision situation in the experiment using an exemplary decision screen. Both treatment options Treatment A and Treatment B are simply depicted in an abstract manner, without indicating actual medical services. The selection of a treatment option (either Treatment A or Treatment B) determines the profit of the physician as well as the benefit for the patient.

	A	B
Your profit (in €)	7	14
Patient's benefit (in €)	10	2

Your decision:

☐

Treatment A

☐

Treatment B

The interpretation of the screen above is as follows: Suppose a participant has chosen Treatment A. Then the earnings of the participant amounts to 7 and the benefit for the patient amounts to 10. If a participant decides on Treatment B, her earnings (in €) amount to 14 and the benefit for the patient (in €) amounts to 2. You choose your preferred treatment option (either Treatment A or Treatment B) by clicking on it.

☐ I have read and understood the information regarding payment.

D.1.2 Summary of the survey items

Table D.1.1: Description of survey items

Variable	Description	Scale
Economic preferences (Falk et al., 2018, 2016)		
<i>Risk aversion</i>	Equality weighted sum of quantitative and qualitative measure and transformed such that $= 0$ implies risk neutrality, > 0 risk aversion, and < 0 risk seeking.	$[-0.5, 0.5]$
Quantitative item	Switching point in multiple price list (31 hypothetical choices between a lottery and a safe option), rescaled to $[0, 1]$.	
Qualitative item	Self-assessed risk taking on a scale from 0 (“not at all willing to take risks”) to 10 (“very willing to take risks”), rescaled to $[0, 1]$.	
<i>Time discounting</i>	Equality weighted sum of quantitative and qualitative measure and transformed such that $= 0$ implies patience, and > 0 impatience.	$[0, 1]$
Quantitative item	Switching point in a list of 25 hypothetical choices between an early payment “today” and a delayed payment “in 12 months”, rescaled to $[0, 1]$.	
Qualitative item	Self-assessed patience on a scale from 0 (“not at all willing to give up something today [in order to benefit from that in the future]”) to 10 (“very willing to give up something”), rescaled to $[0, 1]$.	
<i>Trust</i>	Equality weighted sum of quantitative and qualitative measure.	$[0, 1]$
Quantitative item	First mover behavior in hypothetical investment game, rescaled to $[0, 1]$.	
Qualitative item	Self-assessment on how much the statement “As long as I am not convinced otherwise, I assume that people have only the best intentions” describe subjects on a scale from 0 (“does not describe me at all”) to 10 (“describes me very well”), rescaled to $[0, 1]$.	
<i>Altruism</i>	Equality weighted sum of quantitative and qualitative measure.	$[0, 1]$
Quantitative item	Donation to a good cause after hypothetical lottery win of 1000 Euro, rescaled to $[0, 1]$.	
Qualitative item	Self-assessed altruism on a scale from 0 (“not at all willing to share [with others without expecting anything in return when it comes to a good cause]”) to 10 (“very willing to share”), rescaled to $[0, 1]$.	
<i>Positive reciprocity</i>	Equality weighted sum of two quantitative measures.	$[0, 1]$
Quantitative item 1	Second mover behavior in hypothetical investment game (average of four scenarios), rescaled to $[0, 1]$.	
Quantitative item 2	Size of thank-you gift following hypothetical scenario, rescaled to $[0, 1]$.	
<i>Negative reciprocity</i>	Equality weighted sum of quantitative and qualitative measure.	$[0, 1]$
Quantitative item	Minimum acceptable offer in hypothetical ultimatum game, rescaled to $[0, 1]$.	
Qualitative item	Self-assessed negative reciprocity on a scale from 0 (“not at all willing to punish [unfair behavior even if it is costly]”) to 10 (“very willing to punish”), rescaled to $[0, 1]$.	
Personality traits		
	Self-assessed description of personality according to following items (either Big Five Inventory by Rammstedt and John (2007), Gosling et al. (2003) or HEXACO Personality Inventory by Ashton and Lee (2009). For each trait, we calculated the average score of the respective items and rescaled it to $[-1, 1]$.	
<i>Openness(to experience)</i>		$[-1, 1]$
Big Five	On a scale from 1[Disagree strongly] to 7[Agree strongly]: “I see myself as someone who has few artistic interests.(reverse code)”, “I see myself as someone who has an active imagination.”	

Continued on next page

Variable	Description	Scale
HEXACO	On a scale from 1[Disagree strongly] to 5[Agree strongly]: "I would be quite bored by a visit to an art gallery.(reversed code)", "I'm interested in learning about the history and politics of other countries.", "I would enjoy creating a work of art, such as a novel, a song, or a painting.", "I think that paying attention to radical ideas is a waste of time.(reverse code)", "If I had the opportunity, I would like to attend a classical music concert.", "I've never really enjoyed looking through an encyclopedia.(reverse code)", "People have often told me that I have a good imagination.", "I like people who have unconventional views.", "I don't think of myself as the artistic or creative type. (reverse code)", "I find it boring to discuss philosophy.(reverse code)".	
<i>Conscientiousness</i>		[-1, 1]
Big Five	On a scale from 1[Disagree strongly] to 7[Agree strongly]: "I see myself as someone who tends to be lazy.(reverse code)", "I see myself as someone who does a thorough job."	
HEXACO	On a scale from 1[Disagree strongly] to 5[Agree strongly]: "I plan ahead and organize things, to avoid scrambling at the last minute.", "I often push myself very hard when trying to achieve a goal.", "When working on something, I don't pay much attention to small details.(reverse code)", "I make decisions based on the feeling of the moment rather than on careful thought.(reverse code)", "When working, I sometimes have difficulties due to being disorganized.(reverse code)", "I do only the minimum amount of work needed to get by.(reverse code)", "I always try to be accurate in my work, even at the expense of time.", "I make a lot of mistakes because I don't think before I act.(reverse code)", "People often call me a perfectionist.", "I prefer to do whatever comes to mind, rather than stick to a plan.(reverse code)".	
<i>Agreeableness (versus Anger)</i>		[-1, 1]
Big Five	On a scale from 1[Disagree strongly] to 7[Agree strongly]: "I see myself as someone who tends to find fault with others.(reverse code)", "I see myself as someone who is generally trusting.", "I see myself as someone who is considerate and kind to almost everyone.(additional)"; rescaled to [-1, 1].	
HEXACO	On a scale from 1[Disagree strongly] to 5[Agree strongly]: "I rarely hold a grudge, even against people who have badly wronged me.", "People sometimes tell me that I am too critical of others.(reverse code)", "People sometimes tell me that I'm too stubborn.(reverse code)", "People think of me as someone who has a quick temper.(reverse code)", "My attitude toward people who have treated me badly is áforgive and forgetá.", "I tend to be lenient in judging other people.", "I am usually quite flexible in my opinions when people disagree with me.", "Most people tend to get angry more quickly than I do.", "Even when people make a lot of mistakes, I rarely say anything negative.", "When people tell me that I'm wrong, my first reaction is to argue with them.(reverse code)".	
<i>Extraversion</i>		[-1, 1]
Big Five	On a scale from 1[Disagree strongly] to 7[Agree strongly]: "I see myself as someone who is reserved.(reverse code)", "I see myself as someone who is outgoing, sociable."	
HEXACO	On a scale from 1[Disagree strongly] to 5[Agree strongly]: "I feel reasonably satisfied with myself overall.", "I rarely express my opinions in group meetings.(reverse code)", "I prefer jobs that involve active social interaction to those that involve working alone.", "On most days, I feel cheerful and optimistic.", "I feel that I am an unpopular person.(reverse code)", "In social situations, I'm usually the one who makes the first move.", "The first thing that I always do in a new place is to make friends.", "Most people are more upbeat and dynamic than I generally am.(reverse code)", "I sometimes feel that I am a worthless person.(reverse code)", "When I'm in a group of people, I'm often the one who speaks on behalf of the group.(reverse code)".	
<i>Neuroticism/emotionality</i>		[-1, 1]
Big Five	On a scale from 1[Disagree strongly] to 7[Agree strongly]: "I see myself as someone who is relaxed, handles stress well.(reverse code)", "I see myself as someone who gets nervous easily."	
<i>Continued on next page</i>		

Variable	Description	Scale
HEXACO	On a scale from 1[Disagree strongly] to 5[Agree strongly]: "I would feel afraid if I had to travel in bad weather conditions.", "I sometimes can't help worrying about little things.", "When I suffer from a painful experience, I need someone to make me feel comfortable.", "I feel like crying when I see other people crying.", "When it comes to physical danger, I am very fearful.", "I worry a lot less than most people do.(reverse code)", "I can handle difficult situations without needing emotional support from anyone else.(reverse code)", "I feel strong emotions when someone close to me is going away for a long time.", "Even in an emergency I wouldn't feel like panicking.(reverse code)", "I remain unemotional even in situations where most people get very sentimental.(reverse code).	
Occupationally related items		
Expected income	<p>Calculation based on a subject's self-assessed likelihood (p) of having a monthly net income given a full-time employment five years following the completion of specialty training which falls into the following categories:</p> <ul style="list-style-type: none"> i) less than EUR 3,000 net per month ii) between EUR 3,000 and EUR 3,999 net per month iii) between EUR 4,000 and EUR 4,999 net per month iv) between EUR 5,000 and EUR 5,999 net per month v) more than EUR 6,000 Euro per month <p>The expected value of future income per subject was calculated as follows: $EV = p_i \times 2,500 + p_{ii} \times 3,500 + p_{iii} \times 4,500 + p_{iv} \times 5,500 + p_v \times 6,500$</p>	[2,500, 7,000]
Specialty choice	Selection of the first most preferred specialty from the following list: Anesthesia, General medicine, Neurology/Psychiatry, Orthopedics, Radiology/Nuclear medicine, Forensic medicine, Urology, Dentistry and maxillary surgery, Ophthalmology, Surgery, Dermatology, Gynecology, Otorhinolaryngology, Internal medicine, Pediatrics, Laboratory medicine, Other (namely)	

D.2 Illustration of Delta method

We applied the Delta method as follows. Without covariates, estimated parameters are a 3×1 vector:

$$\eta_0 = \begin{pmatrix} \eta_0^a \\ \eta_0^r \\ \eta_0^\mu \end{pmatrix}.$$

They are transformed back to their original scale thanks to the $l = 3 \times 1$ vector of monotonic continuously differentiable functions g :

$$\theta_0 = g(\eta_0)$$

or more explicitly:

$$\theta_0 = \begin{pmatrix} \theta_0^a \\ \theta_0^r \\ \theta_0^\mu \end{pmatrix} = \begin{pmatrix} g^a(\eta_0^a) \\ g^r(\eta_0^r) \\ g^\mu(\eta_0^\mu) \end{pmatrix}.$$

Then, for a given estimated covariance matrix of the model parameters, $\hat{V}(\hat{\eta}_0)$, the covariance matrix of θ_0 , can be estimated according to the Delta method by:

$$\hat{V}(\hat{\theta}_0) = G(\hat{\eta}_0) \hat{V}(\hat{\eta}_0) G'(\hat{\eta}_0),$$

where $G(\hat{\eta}_0)$ is the 3×3 matrix of $\frac{\partial g^i(\eta)}{\partial \eta^j}$ (the typical element in row i and column j of $G(\hat{\eta}_0)$ is $\frac{\partial g^i(\eta_0)}{\partial \eta_0^j}$):

$$G(\hat{\eta}_0) = \begin{pmatrix} \frac{\partial g^a(\eta_0)}{\partial \eta_0^a} & 0 & 0 \\ 0 & \frac{\partial g^a(\eta_0)}{\partial \eta_0^r} & 0 \\ 0 & 0 & \frac{\partial g^\mu(\eta_0)}{\partial \eta_0^\mu} \end{pmatrix}.$$

We now illustrate the Delta method with dummy variables, where we take gender as an example. Values are transformed back to their original scale with the $l = 3(K+1) \times 1$ vector

of monotonic continuously differentiable functions g . For the constant, the transformation is

$$\theta_0 = g(\eta_0).$$

For the dummy variable, the transformation back to the original scale is:

$$\theta_{female} = g(\eta_0 + \eta_{female}) - g(\eta_0).$$

or more explicitly:

$$\begin{pmatrix} \theta_0^a \\ \theta_{female}^a \\ \theta_0^r \\ \theta_{female}^r \\ \theta_0^\mu \\ \theta_{female}^\mu \end{pmatrix} = \begin{pmatrix} g^a(\eta_0^a) \\ g^a(\eta_0^a + \eta_{female}^a) - g^a(\eta_0^a) \\ g(\eta_0^r) \\ g(\eta_0^r + \eta_{female}^r) - g(\eta_0^r) \\ g(\eta_0^\mu) \\ g(\eta_0^\mu + \eta_{female}^\mu) - g(\eta_0^\mu) \end{pmatrix}.$$

$G(\hat{\eta})$ is now a $3(K+1) \times 3$ matrix of $\frac{\partial g^i(\eta)}{\partial \eta^j}$.

$$\begin{pmatrix} \frac{\partial g^a(\eta_0)}{\partial \eta_0^a} & 0 & 0 & 0 & 0 \\ \frac{\partial g^a(\eta_0 + \eta_{female}^a)}{\partial \eta_0^a} - \frac{\partial g^a(\eta_0)}{\partial \eta_0^a} & \frac{\partial g^a(\eta_0 + \eta_{female}^a)}{\partial \eta_{female}^a} & 0 & 0 & 0 \\ 0 & 0 & \frac{\partial g^r(\eta_0)}{\partial \eta_0^r} & 0 & 0 \\ 0 & 0 & \frac{\partial g^r(\eta_0 + \eta_{female}^r)}{\partial \eta_0^r} - \frac{\partial g^r(\eta_0)}{\partial \eta_0^r} & \frac{\partial g^r(\eta_0 + \eta_{female}^r)}{\partial \eta_{female}^r} & 0 \\ 0 & 0 & 0 & 0 & \dots \end{pmatrix}.$$

For variables other than dummy variables, such as the scores from the preference module, we proceeded as follows, taking altruism as an example. Values are transformed back to their original scale thanks to the $l = 3(K+1) \times 1$ vector of monotonic continuously differentiable functions g . For the constant, the transformation is:

$$\theta_0 = g(\eta_0).$$

Furthermore, for altruism, the transformation back to the original scale is:

$$\theta_{altruism} = g(\eta_0 + 0.1 \times \eta_{altruism}) - g(\eta_0)$$

Put more explicitly, one has:

$$\begin{pmatrix} \theta_0^a \\ \theta_{altruism}^a \\ \theta_0^r \\ \theta_{altruism}^r \\ \theta_0^\mu \\ \theta_{altruism}^\mu \end{pmatrix} = \begin{pmatrix} g^a(\eta_0^a) \\ g^a(\eta_0^a + 0.1 \times \eta_{altruism}^a) - g^a(\eta_0^a) \\ g(\eta_0^r) \\ g(\eta_0^r + 0.1 \times \eta_{altruism}^r) - g(\eta_0^r) \\ g(\eta_0^\mu) \\ g(\eta_0^\mu + 0.1 \times \eta_{altruism}^\mu) - g(\eta_0^\mu) \end{pmatrix}.$$

$G(\hat{\eta})$ is now a $3(K+1) \times 3$ matrix of $\frac{\partial g^i(\eta)}{\partial \eta^j}$.

$$\begin{pmatrix} \frac{\partial g^a(\eta_0)}{\partial \eta_0^a} & 0 & 0 & 0 & 0 \\ \frac{\partial g^a(\eta_0 + 0.1 \times \eta_{altruism}^a)}{\partial \eta_0^a} - \frac{\partial g^a(\eta_0)}{\partial \eta_0^a} & \frac{\partial g^a(\eta_0 + 0.1 \times \eta_{altruism}^a)}{\partial \eta_{altruism}^a} & 0 & 0 & 0 \\ 0 & 0 & \frac{\partial g^r(\eta_0)}{\partial \eta_0^r} & 0 & 0 \\ 0 & 0 & \frac{\partial g^r(\eta_0 + 0.1 \times \eta_{altruism}^r)}{\partial \eta_0^r} - \frac{\partial g^r(\eta_0)}{\partial \eta_0^r} & \frac{\partial g^r(\eta_0 + 0.1 \times \eta_{altruism}^r)}{\partial \eta_{altruism}^r} & 0 \\ 0 & 0 & 0 & 0 & \dots \end{pmatrix}.$$

D.3 Descriptive analysis of the control group

Analogous to Section 4.5.1, we now provide descriptive statistics on patient-regarding behavior and individuals' characteristics of our control group of non-medical students ($N=145$) and compare them to our medical student sample, see Table D.3.1.

Table D.3.1: Descriptive statistics of our control group (non-medical students) and our medical student sample

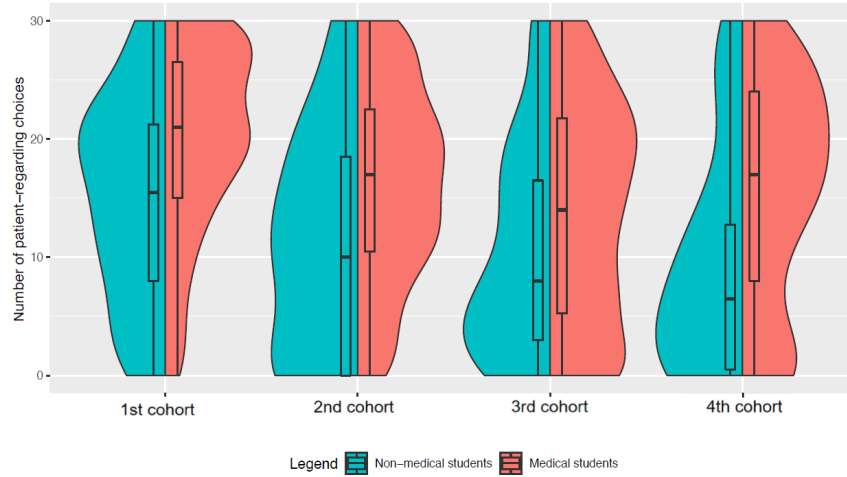
	Non-medical students			Medical students			Δ , t -test
	Mean	s.d.	N	Mean	s.d.	N	p -value
A. Patient-regarding choices							
Total sample	11.9	9.30	145	16.9	9.0	733	<0.001
1ST cohort	15.2	9.0	40	19.8	8.1	259	<0.001
2ND cohort	10.6	9.6	23	16.1	8.3	235	0.001
3RD cohort	11.1	9.1	56	13.8	9.6	158	0.035
4TH cohort	9.7	9.1	26	15.6	9.6	81	0.004
B. Characteristics							
<i>Economic Preferences</i>							
Altruism	0.37	0.14	144	0.38	0.17	729	0.229
Trust	0.50	0.21	144	0.57	0.24	729	<0.001
Positive reciprocity	0.42	0.16	144	0.36	0.18	729	0.017
Negative reciprocity	0.46	0.15	144	0.47	0.16	729	0.439
Risk aversion	0.09	0.15	145	0.07	0.15	731	<0.050
Time discounting	0.31	0.16	145	0.27	0.16	731	0.003
<i>Personality traits</i>							
Extraversion	0.22	0.27	144	0.25	0.41	729	0.271
Agreeableness	0.10	0.27	144	0.09	0.37	729	0.3071
Conscientiousness	0.29	0.28	144	0.39	0.35	729	<0.001
Openness	0.21	0.31	144	0.27	0.44	729	0.055
Emotionality/Neuroticism	0.17	0.31	144	-0.08	0.43	729	<0.001

Notes. This table presents summary statistics on the number of patient-regarding choices and subject's characteristics for our non-medical students sample and compares it to our medical student sample. Comparisons are based on a one-sided t -test with p -values reported in the last column. Subject's characteristics comprise social and economic preferences by the Preference Survey Module (Falk et al., 2016) and personality traits by the 60-item questionnaire of the HEXACO Personality Inventory (Ashton and Lee, 2009). Altruism, trust, positive and negative reciprocity are measured on a scale of [0, 1]. Risk aversion is transformed such that it =0 implies risk neutrality, >0 implies risk aversion and <0 risk seeking. While time discounting = 0 implies patience and a value >0 inpatience. All personality traits are scaled between -1 and 1. Table D.1.1 in Appendix D.1.2 gives a full description of all variables.

Overall, non-medical students make, on average, 11.9 patient-regarding choices. This corresponds to significantly lower patient-regarding behavior by about one third fewer patient-regarding choices. The significance of less patient-regarding choices in non-medical than medical students is prevalent for each cohort comparisons ($p < 0.05$, t -test). Recall

that the progress stages of non-medical students have been synchronized with the medical studies. Students in their first term correspond to freshmen (“1ST cohort”). Students in their second to fourth terms correspond to the pre-clinical phase (“2ND cohort”), students between their fifth to tenth term correspond to the clinical phase (“3RD cohort”), and terms above 10 reflect the practical year (“4TH cohort”). When differentiating between study cohorts for non-medical students, students of the first cohort are the most patient-regarding (15.2 *prcs*), students of the fourth cohort are the least patient-regarding (9.7 *prcs*). While, overall speaking, we find patient-regarding behavior to decrease with study progress, students of the third cohort make slightly more patient-regarding choices than students of second cohort (11.1 compared to 10.6 *prcs*). However, this difference is not statistical significant. Comparing the distributions, the Kolmogorov-Smirnoff-test only rejects the hypothesis of identical distributions for the comparison of *prc* the first and fourth cohort ($p < 0.05$).

Figure D.3.1: Distributions of patient-regarding choices by cohorts for medical and non-medical students



Notes. This figure shows the distributions of the relative frequencies of patient-regarding choices (*prc*) differentiated by study cohorts for medical and non-medical students.

For an illustration of the distributions of *prcs* by study cohorts of our control group, see Figure D.3.1. Students who made no *prc* (referred to as pure profit-maximizers), amount to 5.0% in the first cohort, 30.4% in the second cohort, 12.5% in the third cohort and 23.1% in fourth cohort. Non-medical students who always made the *prc* make up for 5.0% in the

first cohort, 8.7% in the second cohort, 7.1% in the third cohort and 3.9% in fourth cohort. While the overall share of pure patient-regarding altruists is similar in non-medical and medical students (6.8% and 7.8%, respectively), there are percentage-wise more than twice as much pure profit-maximizer among non-medical than medical students (15.2% compared to 6.8%).

While non-medical students are significantly less patient-regarding than medical students, they do not significantly differ in their general altruistic behavior elicited by the Preference Survey Module by Falk et al. (2016, 2018); see Panel B of Table D.3.1. They also do not differ in their preferences for negative reciprocity. Whereas, a comparisons of the other social and economic preferences reveals that our non-medical student sample has lower social preferences for trust and positive reciprocity, but higher preferences for time discounting and risk aversion ($p < 0.05$, t -test). In terms of personality traits, non-medical students appear to be less conscientious, less open but more emotional/neurotic than medical students ($p < 0.10$, t -test).

D.4 Further estimations on CES preferences

D.4.1 Aggregate estimations

Figure D.4.1 shows the distribution of observed heterogeneity implied by the different set of covariates. Table D.4.1 shows the results of the aggregate estimations for the CES preference functional with various sets of covariates. Standard errors are clustered at the individual level. Table D.4.2 shows the corresponding parameter estimates when values are transformed back to the original scale.

Figure D.4.1: Distributions of parameters a , r and noise for the aggregate model with different sets of covariates, CES preferences

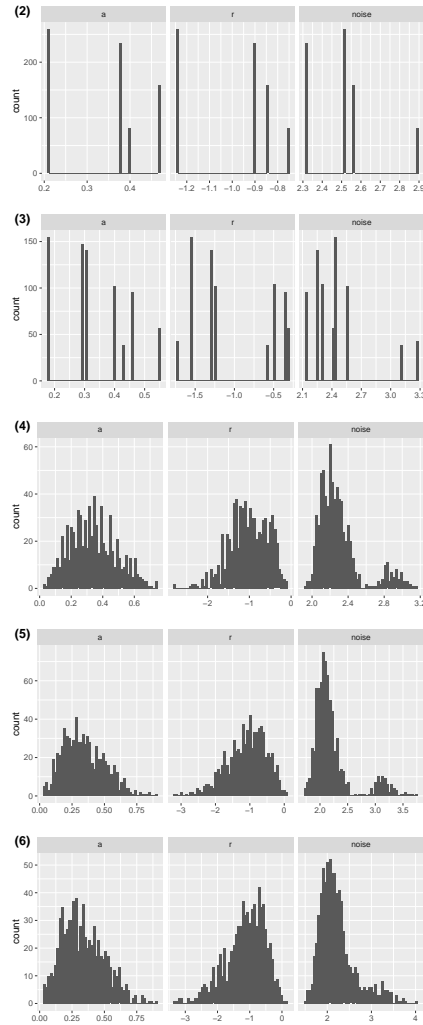


Table D.4.1: Aggregate estimations, parameter estimates, CES preferences

Model	(1)	(2)	(3)	(4)	(5)	(6)
α						
Constant	-0.669*** (0.073)	-1.330*** (0.148)	-0.879*** (0.168)	0.125* (0.208)	0.885*** (0.434)	1.039*** (0.565)
Pre-clinical		0.830*** (0.181)	0.705*** (0.189)	0.653*** (0.196)	0.519*** (0.203)	0.576*** (0.292)
Clinical		1.197*** (0.190)	1.083*** (0.196)	0.942*** (0.199)	0.701*** (0.205)	0.800*** (0.275)
Practical Year		0.917*** (0.259)	0.604*** (0.355)	0.694*** (0.310)	0.394*** (0.375)	0.475*** (0.438)
Female			-0.618*** (0.124)	-0.257*** (0.125)	-0.192*** (0.120)	-0.217*** (0.193)
General altruism				-3.186*** (0.420)	-2.588*** (0.470)	-2.657*** (0.498)
Risk aversion					-0.815*** (0.560)	-0.688*** (0.585)
Time discounting					0.529*** (0.373)	0.414*** (0.368)
Trust					-0.861*** (0.277)	-0.944*** (0.318)
Negative reciprocity					0.157 (0.501)	0.103 (0.781)
Positive reciprocity					-1.211*** (0.591)	-1.243*** (0.546)
Emotionality						0.065 (0.225)
Extraversion						0.148*** (0.229)
Agreeableness						0.176*** (0.309)
Conscientiousness						-0.313*** (0.343)
Openness						-0.015 (0.165)
τ						
Constant	-0.671*** (0.050)	-0.807*** (0.061)	-0.401*** (0.112)	-0.303*** (0.145)	-0.395*** (0.285)	-0.310*** (0.358)
Pre-clinical		0.165* (0.108)	0.105* (0.132)	0.103** (0.126)	0.130*** (0.144)	0.164*** (0.192)
Clinical		0.193* (0.142)	0.131 (0.175)	0.208*** (0.147)	0.206*** (0.164)	0.279*** (0.187)
Practical Year		0.245* (0.208)	-0.060 (0.303)	0.077 (0.268)	-0.076 (0.343)	-0.010 (0.379)
Female			-0.539*** (0.097)	-0.309*** (0.102)	-0.285*** (0.107)	-0.271*** (0.146)
General altruism				-0.718*** (0.240)	-0.770*** (0.307)	-0.744*** (0.350)
Risk aversion					-0.280** (0.432)	-0.117 (0.454)
Time discounting					-0.639*** (0.407)	-0.701*** (0.416)
Trust					0.397*** (0.223)	0.368*** (0.262)
Negative reciprocity					0.460*** (0.372)	0.367*** (0.602)
Positive reciprocity					-0.376*** (0.436)	-0.425*** (0.405)
Emotionality						-0.041 (0.161)
Extraversion						0.045 (0.185)
Agreeableness						0.078 (0.223)
Conscientiousness						-0.238*** (0.272)
Openness						0.036 (0.127)
μ						
Constant	0.964*** (0.035)	0.924*** (0.051)	0.838*** (0.073)	0.810*** (0.137)	0.697*** (0.243)	0.668*** (0.270)
Pre-clinical		-0.082 (0.081)	-0.077 (0.099)	-0.090** (0.101)	-0.077* (0.106)	-0.152*** (0.144)
Clinical		0.016 (0.104)	0.047 (0.128)	-0.048 (0.130)	-0.038 (0.133)	-0.084* (0.146)
Practical Year		0.138 (0.128)	0.293*** (0.216)	0.220*** (0.212)	0.388*** (0.238)	0.332*** (0.263)
Female			0.053 (0.083)	-0.059 (0.091)	-0.033 (0.090)	-0.003 (0.111)
General altruism				0.181* (0.229)	0.130 (0.268)	0.087 (0.281)
Risk aversion					-0.086 (0.314)	-0.212* (0.365)
Time discounting					0.137 (0.304)	0.163 (0.342)
Trust					-0.056 (0.207)	-0.057 (0.218)
Negative reciprocity					-0.157 (0.283)	-0.047 (0.381)
Positive reciprocity					0.267** (0.324)	0.353*** (0.340)
Emotionality						-0.037 (0.110)
Extraversion						-0.179*** (0.129)
Agreeableness						0.019 (0.185)
Conscientiousness						0.065 (0.157)
Openness						-0.023 (0.106)
N	733	733	733	729	729	729
Log-likelihood	-13,331.09	-13,002.64	-12,884.71	-12,394.66	-12,028.09	-11,997.39

Notes.

* p<0.10; ** p<0.05; *** p<0.01

Table D.4.2: Aggregate estimations, preference parameters, noise, and marginal effects, CES preferences

Model	(1)	(2)	(3)	(4)	(5)	(6)
α						
Constant	0.339*** (0.016)	0.209*** (0.028)	0.293*** (0.039)	0.531* (0.053)	0.708*** (0.082)	0.776*** (0.115)
Pre-clinical		0.168*** (0.038)	0.163*** (0.045)	0.154*** (0.048)	0.095*** (0.044)	0.086*** (0.056)
Clinical		0.258*** (0.042)	0.257*** (0.046)	0.213*** (0.047)	0.122*** (0.046)	0.109*** (0.061)
Practical year		0.189*** (0.062)	0.138*** (0.087)	0.163*** (0.068)	0.075*** (0.064)	0.086*** (0.057)
Female			-0.111*** (0.025)	-0.064*** (0.031)	-0.041*** (0.027)	-0.039*** (0.042)
General altruism				-0.079*** (0.011)	-0.056*** (0.013)	-0.049*** (0.022)
Risk aversion					-0.017*** (0.010)	-0.011*** (0.012)
Time discounting					0.011*** (0.008)	0.006*** (0.008)
Trust					-0.018*** (0.006)	-0.018*** (0.007)
Negative reciprocity					0.003 (0.011)	-0.001 (0.014)
Positive reciprocity					-0.026*** (0.011)	-0.025*** (0.011)
Emotionality						0.003* (0.008)
Extraversion						0.004** (0.010)
Agreeableness						0.010*** (0.010)
Conscientiousness						-0.016*** (0.011)
Openness						-0.001 (0.007)
r						
Constant	-0.956*** (0.097)	-1.240*** (0.160)	-0.493*** (0.195)	-0.354*** (0.199)	-0.485*** (0.391)	-0.155 (0.490)
Pre-clinical		0.342* (0.240)	0.149 (0.208)	0.132** (0.170)	0.181*** (0.204)	0.181*** (0.218)
Clinical		0.393* (0.286)	0.183 (0.256)	0.254*** (0.189)	0.276*** (0.229)	0.277*** (0.254)
Practical year		0.486* (0.408)	-0.092 (0.501)	0.100 (0.335)	-0.118 (0.532)	0.246** (0.309)
Female			-1.066*** (0.210)	-0.490*** (0.188)	-0.489*** (0.235)	-0.370*** (0.272)
General altruism				-0.101*** (0.030)	-0.119*** (0.061)	-0.084*** (0.078)
Risk aversion					-0.042** (0.058)	-0.0004 (0.059)
Time discounting					-0.098*** (0.063)	-0.097*** (0.078)
Trust					0.058*** (0.039)	0.029*** (0.037)
Negative reciprocity					0.067*** (0.063)	0.027* (0.080)
Positive reciprocity					-0.057*** (0.061)	-0.054*** (0.054)
Emotionality						-0.003 (0.038)
Extraversion						0.004 (0.053)
Agreeableness						0.041*** (0.051)
Conscientiousness						-0.080*** (0.060)
Openness						0.010 (0.034)
μ						
Constant	2.623*** (0.092)	2.519*** (0.130)	2.313*** (0.181)	2.247*** (0.308)	2.007*** (0.504)	1.773*** (0.480)
Pre-clinical		-0.198 (0.197)	-0.171 (0.220)	-0.194** (0.225)	-0.149* (0.208)	-0.266*** (0.225)
Clinical		0.041 (0.266)	0.112 (0.306)	-0.106 (0.287)	-0.075 (0.264)	-0.137 (0.259)
Practical year		0.373 (0.362)	0.787*** (0.657)	0.552*** (0.571)	0.951*** (0.683)	0.293* (0.569)
Female			0.127 (0.206)	-0.128 (0.200)	-0.064 (0.175)	0.004 (0.213)
General altruism				0.041* (0.048)	0.026 (0.053)	0.016 (0.052)
Risk aversion					-0.017 (0.061)	-0.038* (0.063)
Time discounting					0.028 (0.061)	0.042** (0.070)
Trust					-0.011 (0.043)	0.004 (0.038)
Negative reciprocity					-0.031 (0.061)	-0.002 (0.072)
Positive reciprocity					0.054** (0.063)	0.074*** (0.062)
Emotionality						-0.011 (0.040)
Extraversion						-0.060*** (0.054)
Agreeableness						-0.020 (0.073)
Conscientiousness						0.034* (0.050)
Openness						-0.018 (0.039)
N	733	733	733	729	729	705
Log-likelihood	-13,331.09	-13,002.64	-12,884.71	-12,394.66	-12,028.09	-11,537.7
Notes.	* p<0.10; ** p<0.05; *** p<0.01					

D.4.2 Finite mixture model

As stated in Section 4.5.3, our econometric model can be extended in several other directions to account for individual heterogeneity. One potential direction to account for heterogeneity identifies distinct preference (and noise) types with a finite mixture model without covariates. We next describe the estimation of distinct preference types using a finite mixture model. Here the population is composed by a finite number of preference types C . Each type $c = 1, \dots, C$ is present in proportion π_c in the population and is characterized by a (distinct) set of parameters θ_c . Each individual's type is a priori unknown and has a probability π_c to belong to type c . The likelihood associated with a choice sequence T for decision-maker i is now written:

$$P_i(\pi_1, \dots, \pi_{C-1}, \theta_1, \dots, \theta_C) = \sum_{c=1}^C \pi_c P_i(\theta_c)$$

Taking account theoretical restrictions on parameters (with $\theta_c = g(\zeta_c)$, the likelihood of the sequence of choices for subject i can be written:

$$P_i(\pi_1, \dots, \pi_{C-1}, \zeta_1, \dots, \zeta_C) = \sum_{c=1}^C \pi_c P_i(g(\zeta_c))$$

And the (grand) log-likelihood to be maximized with respect to the set of proportions and parameters $(\pi_1, \dots, \pi_{C-1}, \zeta_1, \dots, \zeta_C)$ is:

$$LL(\pi_1, \dots, \pi_{C-1}, \zeta_1, \dots, \zeta_C) = \sum_i \ln(P_i(\pi_1, \dots, \pi_{C-1}, \zeta_1, \dots, \zeta_C))$$

The log-likelihood is known to be difficult to maximize directly. We follow Bruhin et al. (2019) and first maximize the log-likelihood with an Expectation-Maximisation procedure and then move to a direct maximisation.

Subjects can be classified ex-post in the preference type that best characterize their behavior based on parameters estimates $(\pi_1, \dots, \pi_{C-1}, \theta_1, \dots, \theta_C)$. For a given set of proportions and parameters, the probability for subject i associated with a choice sequence T to belong to

type c is:

$$\tau_{ic} = \frac{\pi_c P_i(\theta_c)}{\sum_{k=1}^C \pi_k P_i(\theta_k)}$$

When evaluated with proportions and parameter estimates at maximum likelihood, τ_{ic} are ex-post probabilities to belong to type c for subject i .

In order to assess the superiority of a finite mixture model over a basic aggregate model with no heterogeneity, we compute several criteria for model selection. First, we compute the well-known Akaike Information criterion (AIC). For a model with C types, the AIC is defined as:

$$AIC(C) = -2LL(\pi_1, \dots, \pi_{C-1}, \zeta_1, \dots, \zeta_C) + 2n_m$$

where n_m denotes the number of estimated parameters. The higher the number of parameters to be estimated, the larger the penalty included in the criterion. The preferred model is the one with the minimum AIC value. We also compute the Bayesian information criterion (BIC):

$$BIC(C) = -2LL(\pi_1, \dots, \pi_{C-1}, \zeta_1, \dots, \zeta_C) + n \ln(N)$$

Similarly, the preferred model is the one with the minimum BIC value.

The last two criteria are based on entropy $E(C)$:

$$E(C) = - \sum_{c=1}^C \sum_{i=1}^N \tau_{ic} \ln(\tau_{ic})$$

where τ_{ic} are ex-post probabilities to belong to type c for subject i . If each subject is unambiguously classified in a given type, then the entropy is equal to 0. On the opposite, when ex-post probabilities to belong to each type are away from 0 or 1, the entropy is large. The third criteria, the normalized entropy criterion (NEC) is equal to entropy normalized by the difference in log-likelihood between a model with C types and the aggregate model:

$$NEC(C) = \frac{E(C)}{LL(\pi_1, \dots, \pi_{C-1}, \zeta_1, \dots, \zeta_C) - LL(\zeta)}$$

The preferred model is the one with the minimum NEC value.

The fourth criteria is the integrated completed likelihood criterion (ICL). This criteria corresponds to the BIC criteria with entropy as included as additional penalty term. Like the NEC, ICL favors C values giving rise to partitioning the data with the greatest evidence. For $C = 1$, the NEC criterion is, by definition, not defined and ICL is equal to BIC.

Estimation results. We estimate a series of finite mixture model in the fashion of Bruhin et al. (2019) to characterize the heterogeneity of preferences in the subject population. We follow Bruhin et al. (2019) and estimate models with $C = 2, 3, 4$ types and then compare these models to the aggregate model without covariates. We use the AIC, BIC, NEC and ICL criteria to assess the best model(s) in terms of fitting.

Table D.4.3 shows the values of criteria for CES preference functional. According to the NEC criterion a model with two types performs best. According to the other criteria, the model with four types is the best model. The latter suggests that heterogeneity in preference types is substantial. This justifies to investigate heterogeneity more thoughtfully with a random coefficient model.

Table D.4.3: Model selection criteria for finite mixture model, CES preferences

C	AIC	BIC	NEC	ICL
1	26668.18	26681.97	NA	26681.97
2	20147.57	20179.75	0.007	20202.78
3	17922.62	17973.19	0.010	18018.37
4	16950.22	17019.18	0.015	17090.13

Table D.4.4 shows the characterization of CES preference types for $C = 2$ and $C = 4$. For $C = 2$, two types emerge: a strongly altruistic type with a low personal share parameter equal to 0.135 and a moderately altruistic type with a high personal share parameter equal to 0.76. For $C = 4$, two additional types appear: extremely altruistic with a share parameter close to zero and Rawlsian type with largely negative r parameter. Both for $C = 2$ and $C = 4$ noise is significantly higher for the most selfish types.

Table D.4.4: Finite mixture model, parameter estimates and preference parameters, CES preferences

	C=2		C=4			
	strongly altruistic	moderately altruistic	Rawls	extremely altr.	strongly altr.	moderately altr.
π	0.609 (0.019)	0.391 (0.019)	0.295 (0.019)	0.207 (0.016)	0.333 (0.020)	0.164 (0.014)
a	0.135 (0.011)	0.756 (0.021)	1 (0.023)	0.036 (0.022)	0.238 (0.011)	0.739 (0.014)
r	-1.522 (0.102)	-1.861 (0.238)	-272.333 (1902.769)	-1.648 (0.575)	-1.309 (0.091)	0.145 (0.107)
μ	1.662 (0.039)	2.012 (0.075)	2.043 (0.070)	1.062 (0.054)	1.114 (0.040)	1.002 (0.070)
Log-likelihood	-10,066.786	NA	NA	-8,460.111	NA	NA

D.4.3 Random coefficient model

We next detail another econometric specification which incorporates observed and unobserved heterogeneity in a random coefficient model. In a random coefficient model, individual parameters are the sum of two components. The first component ηX_i corresponds to observed heterogeneity. It is similar to the estimated parameters in the aggregate model with covariates. The second component corresponds to unobserved heterogeneity. We denote ψ_i , the 3×1 vector representing this second component. Individual parameters are defined as the sum of observed and unobserved heterogeneity:

$$\theta_i = g(\eta X_i + \psi_i)$$

We assume ψ_i follows a multivariate normal distribution independent of the regressors with covariance matrix Ω . For the sake of readability, we abuse notation and denote the components of Ω :

$$\Omega = \begin{pmatrix} \Omega_{[a,a]} & \Omega_{[a,r]} & \Omega_{[a,\mu]} \\ \Omega_{[a,r]} & \Omega_{[r,r]} & \Omega_{[r,\mu]} \\ \Omega_{[a,\mu]} & \Omega_{[r,\mu]} & \Omega_{[\mu,\mu]} \end{pmatrix}$$

For each subject i , $P_i(\theta_i) = P_i(g(\eta X_i + \psi_i))$ is the choice probability conditional on η and ψ_i . Integrating over ψ gives the unconditional probability for subject i (for given η):

$$P_i(\eta, \Omega) = \int_{\mathbb{R}^3} P_i(g(\eta X_i + \psi)) f(\psi|\Omega) d\psi$$

The (grand) log-likelihood is to be maximized with respect to η and Ω is:

$$LL(\eta, \Omega) = \sum_i \ln(P_i(\eta, \Omega))$$

We use Bayesian hierarchical models to estimate this log-likelihood. Corresponding methods are described at length in Train (2009).

Estimation results. Table D.4.5 shows the results from the random coefficient model for the CES preference functional with various sets of covariates. The components of the Ω matrix are given in a separate Table D.4.6. Table D.4.5 shows the results for parameter estimates.

Table D.4.7 shows the estimation results with transformation of the estimated parameters. The share parameter a increases with term, with a maximum attained for clinical studies. This effect remains stable after controlling for gender, altruism, preferences and personality measures. Women and altruist-oriented subjects have lower personal share parameter values. Trust, positive reciprocity and conscientiousness also decrease the estimated value of the personal share parameter. Compare to the aggregate estimations, risk aversion is no longer significant in the random coefficient model. The same applies to agreeableness and extraversion. Still, discounting increases the value of a .

Parameter r also tend to increase with term, but with no firm results among the different estimated models. Estimations show this parameter decrease with gender (being a woman leading to lower values of r and hence higher elasticity of substitution), altruism, discounting and risk aversion. The impact of positive reciprocity and negative reciprocity are no longer significant. On the other hand, trust leads to higher values of r . Noise tend to be lower

for pre-clinical students and larger for students in practical years. Discounting is found to increase noise.

Figure D.4.2 shows the typical indifference curves for the different terms based on preference parameters from model (2). Figure D.4.3 shows the distribution of observed heterogeneity implied by the different set of covariates and Figure D.4.4 shows the corresponding distribution of unobserved heterogeneity.

Table D.4.5: Random coefficient model, parameter estimates, CES preferences

Model	(1)	(2)	(3)	(4)	(5)	(6)
α						
Constant	-0.784*** (0.073)	-1.359*** (0.139)	-0.999*** (0.115)	0.027 (0.175)	0.539** (0.213)	0.806** (0.357)
Pre-clinical		0.678*** (0.187)	0.599*** (0.156)	0.513*** (0.142)	0.434*** (0.098)	0.438*** (0.146)
Clinical		1.165*** (0.173)	1.127*** (0.157)	0.879*** (0.143)	0.743*** (0.123)	0.695*** (0.146)
Practical year		0.912*** (0.229)	0.793*** (0.209)	0.684*** (0.164)	0.586*** (0.176)	0.539*** (0.132)
Female			-0.495*** (0.102)	-0.257** (0.115)	-0.175* (0.099)	-0.133 (0.107)
General altruism				-2.788*** (0.459)	-2.291*** (0.334)	-2.303*** (0.677)
Risk aversion					-0.457 (0.281)	-0.379 (0.273)
Time discounting					0.447** (0.184)	0.262 (0.162)
Trust					-0.816*** (0.209)	-0.785*** (0.205)
Negative reciprocity					0.228 (0.195)	-0.115 (0.220)
Positive reciprocity					-0.882** (0.381)	-0.838*** (0.218)
Emotionality						-0.027 (0.106)
Extraversion						0.001 (0.096)
Agreeableness						0.065 (0.154)
Conscientiousness						-0.245** (0.119)
Openness						-0.070 (0.085)
τ						
Constant	-0.604*** (0.057)	-0.694*** (0.090)	-0.289*** (0.100)	0.007 (0.164)	-0.008 (0.219)	-0.046 (0.216)
Pre-clinical		-0.033 (0.131)	-0.068 (0.131)	-0.118 (0.125)	-0.017 (0.094)	-0.014 (0.128)
Clinical		0.246* (0.135)	0.252 (0.165)	0.152 (0.140)	0.268*** (0.100)	0.264* (0.141)
Practical year		0.385** (0.176)	0.365** (0.186)	0.312* (0.163)	0.383*** (0.137)	0.414*** (0.149)
Female			-0.624*** (0.102)	-0.508*** (0.092)	-0.488*** (0.086)	-0.466*** (0.108)
General altruism				-0.844*** (0.297)	-1.093*** (0.324)	-1.019*** (0.353)
Risk aversion					-0.363** (0.176)	-0.278 (0.229)
Time discounting					-1.215*** (0.269)	-1.267*** (0.228)
Trust					0.466*** (0.165)	0.299 (0.199)
Negative reciprocity					0.246 (0.217)	0.331 (0.253)
Positive reciprocity					0.017 (0.141)	0.314* (0.187)
Emotionality						-0.146 (0.114)
Extraversion						-0.170 (0.117)
Agreeableness						0.165 (0.139)
Conscientiousness						-0.224 (0.141)
Openness						0.122 (0.105)
μ						
Constant	-0.259*** (0.033)	-0.191*** (0.054)	-0.335*** (0.060)	-0.505*** (0.114)	-0.634*** (0.112)	-0.652*** (0.176)
Pre-clinical		-0.007 (0.073)	0.007 (0.078)	0.026 (0.080)	-0.012 (0.060)	-0.001 (0.069)
Clinical		-0.239*** (0.076)	-0.244*** (0.091)	-0.208** (0.097)	-0.266*** (0.078)	-0.229*** (0.078)
Practical year		-0.098 (0.101)	-0.091 (0.098)	-0.068 (0.097)	-0.085 (0.093)	-0.140 (0.112)
Female			0.221*** (0.052)	0.184*** (0.065)	0.177*** (0.056)	0.212*** (0.070)
General altruism				0.453** (0.210)	0.637*** (0.199)	0.520*** (0.157)
Risk aversion					0.007 (0.169)	-0.098 (0.209)
Time discounting					0.829*** (0.177)	0.656*** (0.154)
Trust					-0.176* (0.105)	0.030 (0.108)
Negative reciprocity					-0.088 (0.156)	-0.011 (0.153)
Positive reciprocity					0.018 (0.136)	-0.038 (0.146)
Emotion						0.080 (0.070)
Extraversion						-0.038 (0.080)
Agreeableness						-0.041 (0.079)
Conscientiousness						0.008 (0.093)
Openness						-0.123* (0.063)
N	733	733	733	729	729	705
Log-likelihood	-5,562.34	-5,574.82	-5,555.61	-5,518.31	-5,508.63	-5,311.27

Notes.

*p<0.10; **p<0.05; ***p<0.01

Table D.4.6: Random coefficient model, covariance parameter estimates, CES preferences

Model	(1)	(2)	(3)	(4)	(5)	(6)
$\Omega_{[a,a]}$	2.068*** (0.205)	1.902*** (0.187)	1.808*** (0.173)	1.529*** (0.156)	1.426*** (0.134)	1.362*** (0.128)
$\Omega_{[r,r]}$	1.370*** (0.154)	1.382*** (0.164)	1.251*** (0.141)	1.209*** (0.139)	1.119*** (0.122)	1.149*** (0.131)
$\Omega_{[\mu,\mu]}$	0.365*** (0.040)	0.359*** (0.041)	0.350*** (0.041)	0.343*** (0.043)	0.324*** (0.039)	0.326*** (0.039)
$\Omega_{[a,r]}$	0.488*** (0.127)	0.466*** (0.122)	0.360*** (0.114)	0.276*** (0.101)	0.301*** (0.090)	0.267*** (0.093)
$\Omega_{[a,\mu]}$	-0.225*** (0.066)	-0.192*** (0.062)	-0.151** (0.060)	-0.119** (0.054)	-0.125** (0.052)	-0.135*** (0.052)
$\Omega_{[r,\mu]}$	-0.532*** (0.062)	-0.526*** (0.064)	-0.470*** (0.060)	-0.460*** (0.060)	-0.413*** (0.053)	-0.421*** (0.057)
N	733	733	733	729	729	705
Log-likelihood	-5,562.34	-5,574.82	-5,555.61	-5,518.31	-5,508.63	-5,311.27
Notes.	*p<0.10; **p<0.05; ***p<0.01					

Figure D.4.2: Indifference curves for different terms based on random coefficient model, CES preferences

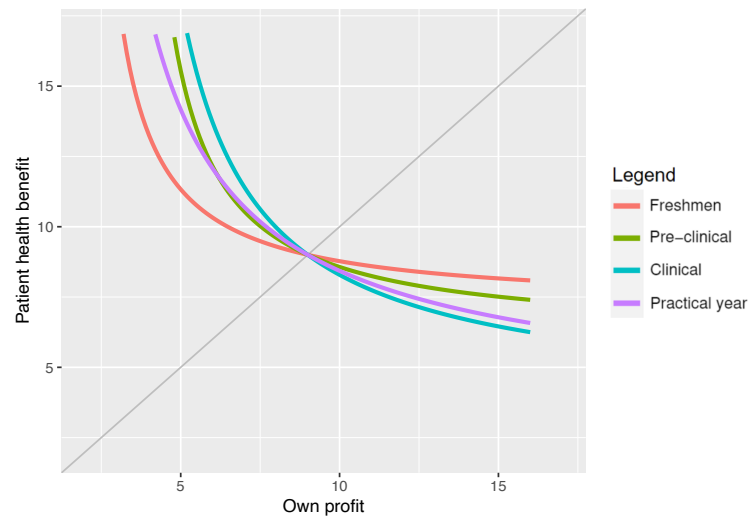


Table D.4.7: Random coefficient model, preference parameters, noise, and marginal effects, CES preferences

Model	(1)	(2)	(3)	(4)	(5)	(6)
α						
Constant	0.314*** (0.016)	0.205*** (0.022)	0.270*** (0.023)	0.507*** (0.043)	0.630*** (0.050)	0.687*** (0.081)
Pre-clinical		0.131*** (0.035)	0.132*** (0.035)	0.123*** (0.032)	0.094*** (0.023)	0.084*** (0.031)
Clinical		0.247*** (0.035)	0.262*** (0.036)	0.203*** (0.029)	0.151*** (0.028)	0.126*** (0.036)
Practical year		0.186*** (0.049)	0.179*** (0.049)	0.161*** (0.033)	0.122*** (0.036)	0.100*** (0.028)
Female			-0.086*** (0.017)	-0.064** (0.029)	-0.041* (0.024)	-0.029 (0.024)
General altruism				-0.069*** (0.012)	-0.054*** (0.008)	-0.049*** (0.013)
Risk aversion					-0.011 (0.007)	-0.008 (0.006)
Time discounting					0.010** (0.005)	0.006 (0.003)
Trust					-0.019*** (0.005)	-0.017*** (0.005)
Negative reciprocity					0.005 (0.005)	-0.002 (0.005)
Positive reciprocity					-0.021** (0.009)	-0.018*** (0.004)
Emotionality						-0.001 (0.004)
Extraversion						0.0001 (0.004)
Agreeableness						0.003 (0.007)
Conscientiousness						-0.010** (0.005)
Openness						-0.003 (0.004)
τ						
Constant	-0.833*** (0.105)	-1.009*** (0.182)	-0.342*** (0.132)	-0.007 (0.168)	-0.033 (0.237)	-0.072 (0.233)
Pre-clinical		-0.024 (0.102)	-0.037 (0.073)	-0.024 (0.026)	-0.002 (0.011)	-0.001 (0.014)
Clinical		0.191* (0.109)	0.135 (0.091)	0.029 (0.028)	0.031** (0.014)	0.027 (0.024)
Practical year		0.295* (0.152)	0.192** (0.097)	0.060* (0.035)	0.043*** (0.016)	0.039 (0.025)
Female			-0.363*** (0.070)	-0.106*** (0.026)	-0.061*** (0.018)	-0.049 (0.032)
General altruism				-0.181** (0.073)	-0.144*** (0.054)	-0.098*** (0.027)
Risk aversion					-0.046* (0.026)	-0.028 (0.022)
Time discounting					-0.162*** (0.046)	-0.135*** (0.051)
Trust					0.052*** (0.019)	0.024 (0.021)
Negative reciprocity					0.029 (0.026)	0.030 (0.026)
Positive reciprocity					0.002 (0.017)	0.032 (0.027)
Emotion						-0.013 (0.011)
Extraversion						-0.016 (0.012)
Agreeableness						0.016 (0.016)
Conscientiousness						-0.023 (0.017)
Openness						0.011 (0.011)
μ						
Constant	0.772*** (0.025)	0.827*** (0.044)	0.717*** (0.043)	0.607*** (0.068)	0.534*** (0.061)	0.529*** (0.098)
Pre-clinical		-0.0002 (0.004)	0.001 (0.006)	0.005 (0.017)	-0.004 (0.022)	0.001 (0.036)
Clinical		-0.012*** (0.004)	-0.017*** (0.006)	-0.043* (0.022)	-0.091*** (0.034)	-0.102** (0.041)
Practical year		-0.005 (0.005)	-0.006 (0.007)	-0.014 (0.021)	-0.030 (0.034)	-0.060 (0.054)
Female			0.017*** (0.005)	0.039** (0.016)	0.063*** (0.023)	0.102** (0.047)
General altruism				0.100* (0.052)	0.248** (0.099)	0.273** (0.135)
Risk aversion					0.003 (0.064)	-0.037 (0.091)
Time discounting					0.321*** (0.104)	0.346** (0.147)
Trust					-0.062 (0.038)	0.018 (0.052)
Negative reciprocity					-0.028 (0.051)	-0.006 (0.073)
Positive reciprocity					0.004 (0.044)	-0.008 (0.058)
Emotion						0.039 (0.036)
Extraversion						-0.017 (0.039)
Agreeableness						-0.019 (0.038)
Conscientiousness						0.002 (0.048)
Openness						-0.058 (0.036)
N	733	733	733	729	729	705
Log-likelihood	-5,562.34	-5,574.82	-5,555.61	-5,518.31	-5,508.63	-5,311.27
Notes.	* p<0.10; ** p<0.05; *** p<0.01					

Figure D.4.3: Distributions of parameters a , r and noise based on observed heterogeneity for the random coefficient model with different sets of covariates, CES preferences

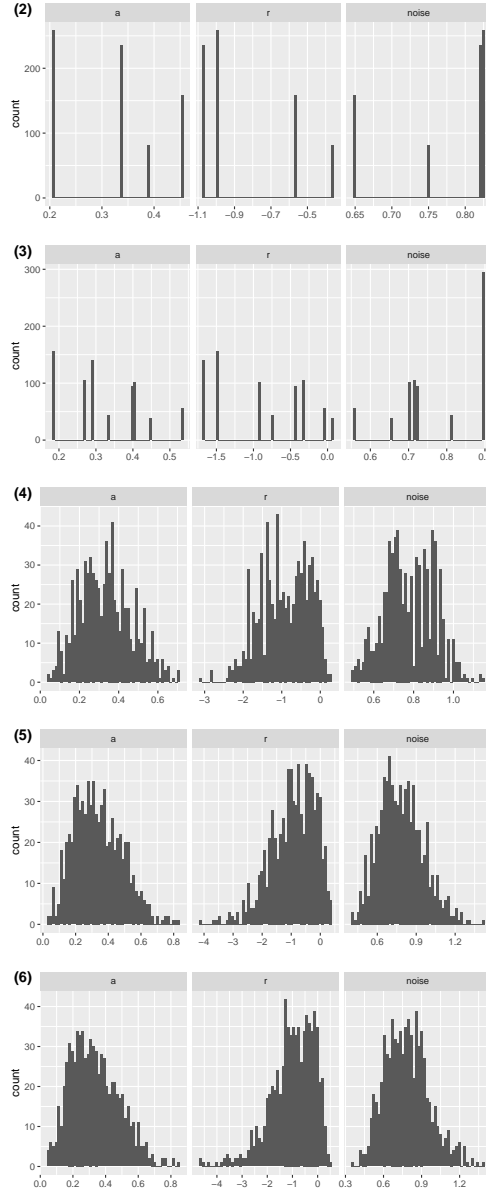
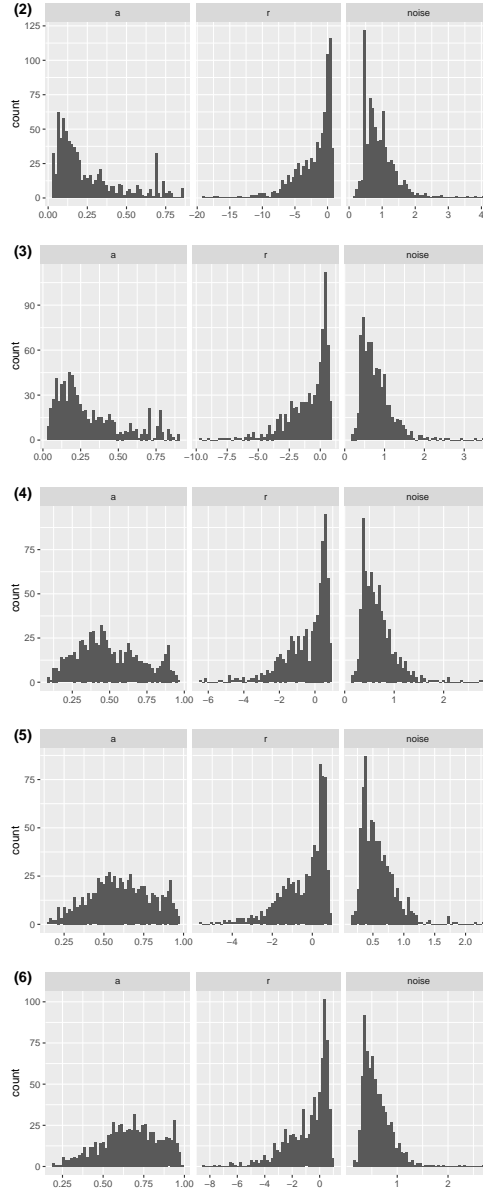


Figure D.4.4: Distributions of parameters a , r and noise based on unobserved heterogeneity for the random coefficient model with different sets of covariates, CES preferences



D.4.4 Individual estimates

A last alternative to account for individual heterogeneity is to estimate the preference and noise parameters for each individual separately. For individual estimations, the individual log-likelihood $\ln(P_i(\theta_i, \mu_i))$ is maximized with respect to θ_i and μ_i for each subject i . Because individual estimations might depend on initial values of parameters (Bruhin et al., 2019), for each individual estimation we run a series of 3^3 estimations with different starting values. The set of starting values is constructed as a grid of all combinations of $a = \{0.1, 0.5, 0.9\}$, $r = \{-10, -2, 0.5\}$ and $\mu = \{0.5, 2, 5\}$ for the CES preferences. The individual estimates corresponding to the best fitting parameter estimates out of the 3^3 possible values. When convergence is not attained, all parameter estimates at the subject level are treated as missing.

Estimation results. Table D.4.8 shows the results of the individual estimations for the CES preferences. Table D.4.8 reports the descriptive statistics for the parameter estimates (Column 1) and the preference estimates transformed back to their original scale (Column 2).

Table D.4.8: Individual results: descriptive statistics, median and interquartile range, CES preferences

	Parameter estimates	Preference parameters
a	-0.61	0.352
Q1-Q3	-1.473 - 0.045	0.186 - 0.511
r	-0.809	-1.245
Q1-Q3	-1.656 - 0.044	-4.239 - 0.043
μ	-0.462	0.63
Q1-Q3	-1.443 - 0.06	0.236 - 1.062
Log-likelihood	-7.763	
Q1-Q3	-11.448 - -4.225	
N	610	

D.4.5 Estimation results for all subjects

We next report the estimation results for all subjects including our control group of non-medical students for direction of our econometric model. First, we account for observed heterogeneity in the aggregate estimation. Second, we account for both observed and unobserved heterogeneity in a random coefficient model. In what follows, we use cohorts instead of study progress to include non-medical students. The cohorts match study progress in the following way: the first cohort corresponds to freshmen (reference category), the second cohort corresponds to pre-clinical phase, the third cohort corresponds to clinical phase and fourth cohort corresponds to practical year.

Estimation results of aggregate estimation. Table D.4.9 shows the result of the aggregate estimations for the CES preference functional with various sets of covariates with standard errors clustered at the individual subject level. Figure D.4.5 shows the typical indifference curves for medical and non-medical students based on parameter estimates from Model (2).

Figure D.4.5: Indifference curves for medical and non medical students for aggregate estimation, CES preferences

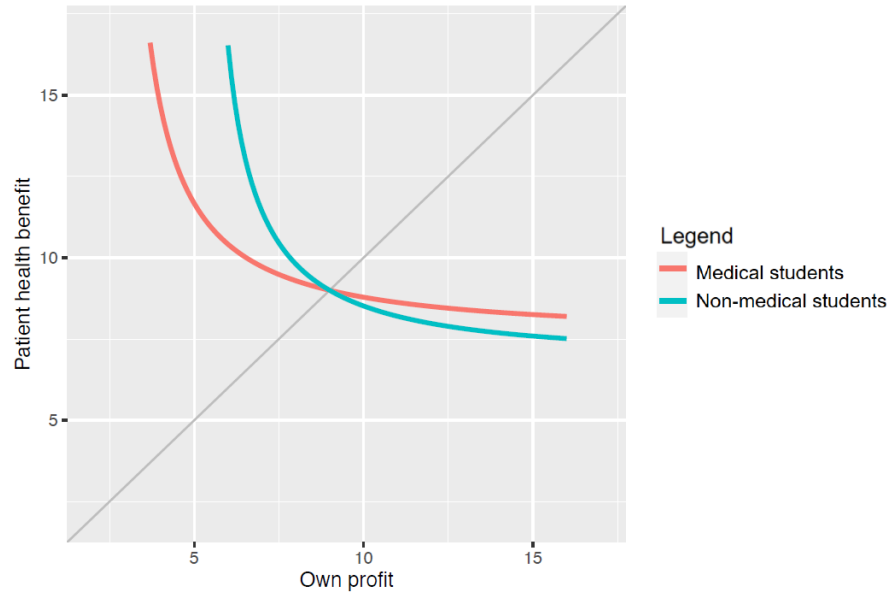


Table D.4.9: Aggregate estimations, parameter estimates for medical and non-medical students, CES preferences

Model	(1)	(2)	(3)	(4)	(5)
<i>a</i>					
Constant	-0.536*** (0.066)	-1.397*** (0.154)	-0.774*** (0.148)	0.086 (0.188)	0.779*** (0.347)
2ND cohort		0.895*** (0.191)	0.582*** (0.179)	0.682*** (0.176)	0.576*** (0.175)
3RD cohort		1.232*** (0.199)	0.961*** (0.181)	0.932*** (0.178)	0.726*** (0.181)
4TH cohort		1.079*** (0.209)	0.659*** (0.256)	0.792*** (0.234)	0.576*** (0.303)
Female			-0.733*** (0.161)	-0.330*** (0.120)	-0.264*** (0.123)
General altruism				-2.989*** (0.421)	-2.455*** (0.461)
Risk aversion					-0.957*** (0.461)
Time discounting					0.551*** (0.360)
Trust					-0.809*** (0.249)
Negative reciprocity					0.025 (0.463)
Positive reciprocity					-0.959*** (0.479)
Non-medical students		0.894*** (0.227)	0.923*** (0.178)	0.844*** (0.193)	0.778*** (0.186)
<i>r</i>					
Constant	-0.670*** (0.051)	-0.844*** (0.066)	-0.318*** (0.132)	-0.330*** (0.138)	-0.393*** (0.250)
2ND cohort		0.201*** (0.133)	-0.010 (0.146)	0.131*** (0.118)	0.181*** (0.127)
3RD cohort		0.121 (0.224)	-0.111 (0.283)	0.136** (0.143)	0.187*** (0.145)
4TH cohort		0.433*** (0.180)	0.007 (0.239)	0.194*** (0.199)	0.051 (0.303)
Female			-0.620*** (0.141)	-0.358*** (0.104)	-0.346*** (0.117)
General altruism				-0.575*** (0.259)	-0.676*** (0.321)
Risk aversion					-0.371*** (0.396)
Time discounting					-0.659*** (0.402)
Trust					0.457*** (0.211)
Negative reciprocity					0.251** (0.375)
Positive reciprocity					-0.241** (0.391)
Non-medical students		-0.203** (0.289)	0.046 (0.237)	-0.100 (0.263)	-0.047 (0.256)
<i>μ</i>					
Constant	1.008*** (0.034)	0.938*** (0.055)	0.788*** (0.098)	0.777*** (0.125)	0.597*** (0.233)
2ND cohort		-0.079 (0.103)	0.042 (0.111)	-0.099** (0.093)	-0.098** (0.100)
3RD cohort		0.068 (0.149)	0.211*** (0.162)	0.016 (0.119)	-0.007 (0.118)
4TH cohort		0.0003 (0.158)	0.245*** (0.194)	0.133** (0.183)	0.310*** (0.220)
Female			0.070* (0.109)	-0.055 (0.090)	-0.014 (0.088)
General altruism				0.250** (0.219)	0.237** (0.253)
Risk aversion					-0.024 (0.303)
Time discounting					0.215** (0.282)
Trust					-0.055 (0.189)
Negative reciprocity					-0.063 (0.290)
Positive reciprocity					0.223** (0.292)
Non-medical students		0.179*** (0.157)	0.061 (0.162)	0.204*** (0.154)	0.248*** (0.153)
<i>N</i>	878	878	878	873	873
Log-likelihood	-16,279.32	-15,661.38	-15,486.70	-14,981.08	-14,620.35
Notes. *p<0.10; **p<0.05; ***p<0.01					

Table D.4.10 shows the estimation results when estimated parameters are transformed back to the original scale. The personal share parameter a increases with term, with a maximum attained for the second term category. This effect remains stable after controlling for gender, altruism, preferences and personality measures. Women and altruist-oriented subjects have lower personal share parameter values while non medical student have high higher share parameter a (for oneself). Risk aversion, trust, positive reciprocity decrease the estimated value of the personal share parameter. On the other hand, discounting increases the value of a .

Parameter r also tend to increase with term, but with no firm results among the different regression. Estimations shows this parameter decreases with gender (being a woman leading to lower values of r), altruism, discounting, positive reciprocity and risk aversion. No stable effect is associated with non-medical students. On the other hand, trust and negative reciprocity lead to higher values of r , i.e. to lower values of the elasticity of substitution. Noise tends to be larger for the fourth cohort. Positive reciprocity, altruism and discounting are found to increase noise. Non-medical students also display a higher level of noise. Figure D.4.6 shows the distribution of observed heterogeneity implied by the different set of covariates.

Estimation results of random coefficient model. Table D.4.11 shows the result from the random coefficient model for the CES preference functional with various sets of covariates. The components of the Ω matrix are given in table D.4.12. Table D.4.13 shows the estimation results when estimated parameters are transformed back to the original scale. The personal share parameter a increases with term, with a maximum attained for the third cohort. This effect remains stable after controlling for gender, altruism and preferences. Women and altruist-oriented subjects have lower personal share parameter values. Trust, risk aversion and positive reciprocity also decrease the estimated value of the personal share parameter. Compared to the aggregate estimation risk aversion is significant and discounting is no longer significant. Non-medical students have higher value of the personal share parameter.

Parameter r also tend to increase with term for the third and fourth cohort. Estimations show this parameter decreases with gender (being a woman leading to lower values of r),

altruism, discounting and risk aversion. The impact of positive reciprocity and negative reciprocity are no longer significant. On the other hand, trust lead to higher values of r . No systematic impact is associated with non-medical students on r . Noise tend to be lower for the third cohort. Discounting and gender (being a women) are found to increase noise. No systematic impact is associated with non-medical students on noise parameter.

Figure D.4.7 shows the typical indifference curves for medical and non-medical students based on parameter estimates for Model (2).

Figure D.4.6: Distributions of parameters a , r and noise for the aggregate model with different sets of covariates for medical and non medical students, CES preferences

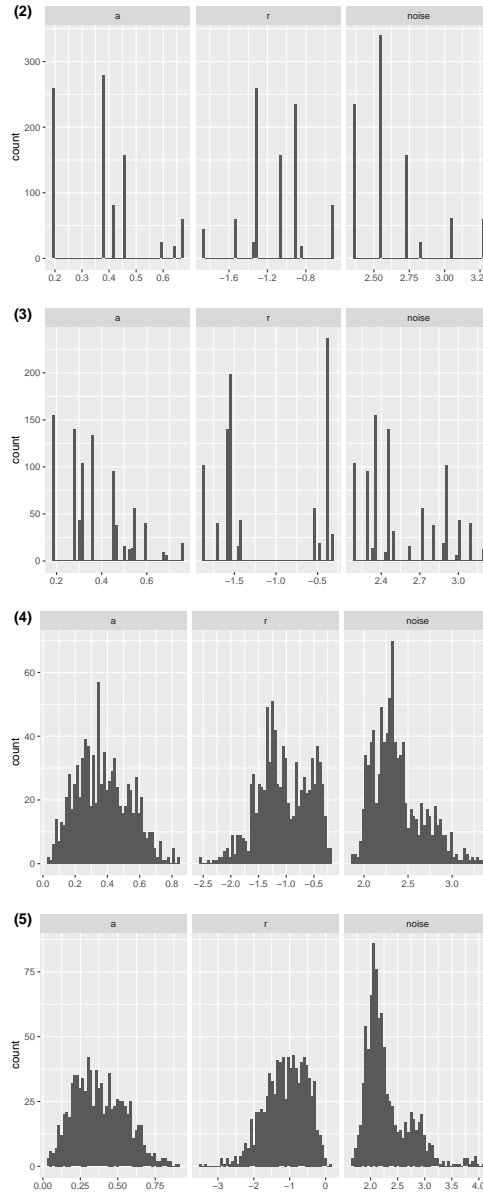


Table D.4.10: Aggregate estimations, preference parameters, noise, and marginal effects for medical and non-medical students, CES preferences

Model	(1)	(2)	(3)	(4)	(5)
<i>a</i>					
Constant	0.369*** (0.015)	0.198*** (0.025)	0.316*** (0.032)	0.522 (0.047)	0.685*** (0.075)
2ND cohort		0.179*** (0.036)	0.137*** (0.041)	0.162*** (0.043)	0.110*** (0.041)
3RD cohort		0.261*** (0.040)	0.231*** (0.042)	0.213*** (0.042)	0.133*** (0.043)
4TH cohort		0.223*** (0.043)	0.156*** (0.061)	0.185*** (0.053)	0.109*** (0.056)
Female			−0.134*** (0.030)	−0.082*** (0.029)	−0.059*** (0.029)
General altruism				−0.075*** (0.011)	−0.055*** (0.011)
Risk aversion					−0.021*** (0.010)
Time discounting					0.012*** (0.008)
Trust					−0.018*** (0.005)
Negative reciprocity					0.001 (0.010)
Positive reciprocity					−0.021*** (0.010)
Non-medical students		0.179*** (0.051)	0.222*** (0.043)	0.196*** (0.041)	0.140*** (0.039)
<i>r</i>					
Constant	−0.955*** (0.100)	−1.326*** (0.154)	−0.374*** (0.181)	−0.391*** (0.192)	−0.482*** (0.370)
2ND cohort		0.424*** (0.270)	−0.013 (0.202)	0.171*** (0.160)	0.245*** (0.191)
3RD cohort		0.266 (0.461)	−0.161 (0.427)	0.177** (0.188)	0.253*** (0.207)
4TH cohort		0.817*** (0.297)	0.009 (0.327)	0.245*** (0.243)	0.074 (0.429)
Female			−1.179*** (0.253)	−0.599*** (0.201)	−0.612*** (0.275)
General altruism				−0.082*** (0.033)	−0.104*** (0.056)
Risk aversion					−0.056*** (0.061)
Time discounting					−0.101*** (0.059)
Trust					0.066*** (0.039)
Negative reciprocity					0.037** (0.058)
Positive reciprocity					−0.036** (0.057)
Non-medical students		−0.524** (0.817)	0.062 (0.312)	−0.146 (0.401)	−0.071 (0.394)
<i>μ</i>					
Constant	2.739*** (0.094)	2.554*** (0.140)	2.199*** (0.215)	2.175*** (0.271)	1.816*** (0.424)
2ND cohort		−0.195 (0.249)	0.094 (0.249)	−0.205** (0.199)	−0.170** (0.183)
3RD cohort		0.180 (0.406)	0.516*** (0.430)	0.035 (0.261)	−0.013 (0.214)
4TH cohort		0.001 (0.403)	0.611*** (0.517)	0.308** (0.441)	0.661*** (0.531)
Female			0.159* (0.244)	−0.117 (0.191)	−0.026 (0.159)
General altruism				0.055** (0.044)	0.043** (0.047)
Risk aversion					−0.004 (0.055)
Time discounting					0.039** (0.050)
Trust					−0.010 (0.035)
Negative reciprocity					−0.011 (0.054)
Positive reciprocity					0.041** (0.052)
Non-medical students		0.501*** (0.470)	0.138 (0.376)	0.491*** (0.405)	0.511*** (0.377)
<i>N</i>	878	878	878	873	873
Log-likelihood	−16,279.32	−15,661.38	−15,486.70	−14,981.08	−14,620.35
Notes.				* p<0.10; ** p<0.05; *** p<0.01	

Table D.4.11: Random coefficient model, parameter estimates for medical and non medical students, CES preferences

Model	(1)	(2)	(3)	(4)	(5)
<i>a</i>					
Constant	-0.673*** (0.068)	-1.317*** (0.091)	-1.017*** (0.162)	0.036 (0.138)	0.424 (0.374)
2ND cohort		0.623*** (0.123)	0.665*** (0.174)	0.572*** (0.135)	0.456*** (0.119)
3RD cohort		1.012*** (0.111)	1.112*** (0.198)	0.908*** (0.115)	0.757*** (0.097)
4TH cohort		0.891*** (0.171)	0.925*** (0.251)	0.742*** (0.178)	0.614*** (0.149)
Female			-0.552*** (0.109)	-0.314*** (0.115)	-0.310*** (0.093)
General altruism				-2.894*** (0.431)	-2.411*** (0.323)
Risk aversion					-0.517*** (0.157)
Time discounting					0.456 (0.309)
Trust					-0.672*** (0.167)
Negative reciprocity					0.115 (0.396)
Positive reciprocity					-0.486** (0.240)
Non-medical students		0.714*** (0.133)	0.711*** (0.142)	0.748*** (0.153)	0.636*** (0.117)
<i>r</i>					
Constant	-0.587*** (0.053)	-0.732*** (0.085)	-0.278*** (0.102)	-0.020 (0.150)	-0.310 (0.329)
2ND cohort		0.019 (0.125)	-0.023 (0.107)	-0.062 (0.119)	0.036 (0.114)
3RD cohort		0.281** (0.116)	0.283** (0.129)	0.243* (0.141)	0.339*** (0.131)
4TH cohort		0.410** (0.187)	0.450*** (0.147)	0.341** (0.171)	0.496*** (0.136)
Female			-0.713*** (0.094)	-0.544*** (0.088)	-0.547*** (0.101)
General altruism				-0.897*** (0.312)	-0.945*** (0.297)
Risk aversion					-0.660** (0.310)
Time discounting					-1.127*** (0.257)
Trust					0.524** (0.265)
Negative reciprocity					0.381 (0.290)
Positive reciprocity					0.265 (0.161)
Non-medical students		0.178 (0.122)	0.175 (0.137)	0.247* (0.149)	0.321* (0.166)
<i>μ</i>					
Constant	-0.262*** (0.030)	-0.177*** (0.047)	-0.331*** (0.062)	-0.513*** (0.079)	-0.483*** (0.173)
2ND cohort		-0.020 (0.066)	-0.003 (0.071)	0.020 (0.064)	-0.022 (0.064)
3RD cohort		-0.256*** (0.066)	-0.256*** (0.075)	-0.238*** (0.076)	-0.302*** (0.077)
4TH cohort		-0.116 (0.104)	-0.137 (0.103)	-0.066 (0.096)	-0.162* (0.084)
Female			0.248*** (0.056)	0.184*** (0.055)	0.183*** (0.052)
General altruism				0.529*** (0.173)	0.585*** (0.199)
Risk aversion					0.061 (0.178)
Time discounting					0.771*** (0.141)
Trust					-0.226* (0.121)
Negative reciprocity					-0.164 (0.173)
Positive reciprocity					-0.047 (0.181)
Non-medical students		-0.007 (0.075)	-0.006 (0.075)	-0.010 (0.080)	-0.045 (0.082)
<i>N</i>	878	878	878	873	873
Log-likelihood	-6,615.2	-6,631.76	-6,624.59	-6,582.66	-6,578.67
<i>Notes.</i>				* p<0.10; ** p<0.05; *** p<0.01	

Table D.4.12: Random coefficient model, covariance parameter estimates for medical and non medical students, CES preferences

Model	(1)	(2)	(3)	(4)	(5)
$\Omega_{[a,a]}$	2.326*** (0.211)	2.045*** (0.180)	1.977*** (0.179)	1.727*** (0.152)	1.624*** (0.145)
$\Omega_{[r,r]}$	1.471*** (0.145)	1.445*** (0.148)	1.332*** (0.147)	1.303*** (0.138)	1.245*** (0.128)
$\Omega_{[\mu,\mu]}$	0.368*** (0.038)	0.358*** (0.038)	0.344*** (0.037)	0.337*** (0.035)	0.316*** (0.036)
$\Omega_{[a,r]}$	0.548*** (0.127)	0.481*** (0.112)	0.373*** (0.110)	0.300*** (0.095)	0.332*** (0.097)
$\Omega_{[a,\mu]}$	-0.228*** (0.061)	-0.187*** (0.060)	-0.143*** (0.055)	-0.111** (0.050)	-0.127*** (0.047)
$\Omega_{[r,\mu]}$	-0.544*** (0.059)	-0.528*** (0.058)	-0.474*** (0.055)	-0.466*** (0.055)	-0.432*** (0.054)
N	878	878	878	873	873
Log-likelihood	-6,615.2	-6,631.76	-6,624.59	-6,582.66	-6,578.67
Notes.	* p<0.10; ** p<0.05; *** p<0.01				

Figure D.4.7: Indifference curves for medical and non medical students based on random coefficient model, CES preferences

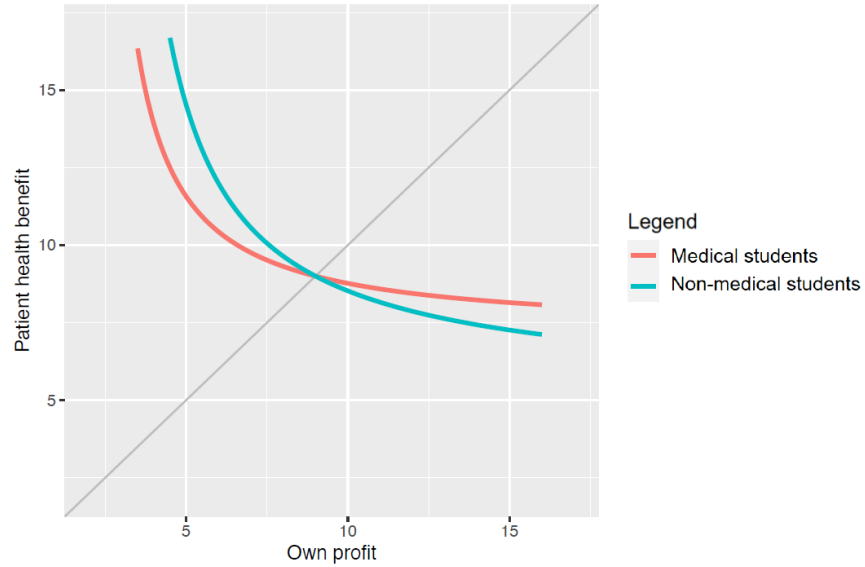


Table D.4.13: Random coefficient model, preference parameters, noise, and marginal effects for medical and non medical students, CES preferences

Model	(1)	(2)	(3)	(4)	(5)
<i>a</i>					
Constant	0.338*** (0.015)	0.212*** (0.015)	0.267*** (0.032)	0.509*** (0.034)	0.601*** (0.086)
2ND cohort		0.122*** (0.025)	0.146*** (0.037)	0.137*** (0.031)	0.099*** (0.024)
3RD cohort		0.213*** (0.025)	0.257*** (0.043)	0.210*** (0.027)	0.157*** (0.027)
4TH cohort		0.184*** (0.038)	0.210*** (0.057)	0.174*** (0.038)	0.129*** (0.031)
Female			-0.093*** (0.022)	-0.077*** (0.028)	-0.074*** (0.024)
General altruism				-0.072*** (0.011)	-0.057*** (0.008)
Risk aversion					-0.012*** (0.004)
Time discounting					0.011 (0.008)
Trust					-0.016*** (0.003)
Negative reciprocity					0.003 (0.009)
Positive reciprocity					-0.011** (0.005)
Non-medical students		0.143*** (0.030)	0.158*** (0.033)	0.176*** (0.032)	0.134*** (0.027)
<i>r</i>					
Constant	-0.802*** (0.095)	-1.087*** (0.177)	-0.328** (0.135)	-0.032 (0.152)	-0.439 (0.477)
2ND cohort		0.015 (0.092)	-0.011 (0.061)	-0.011 (0.023)	0.006 (0.017)
3RD cohort		0.206** (0.087)	0.154** (0.075)	0.047 (0.028)	0.047* (0.025)
4TH cohort		0.297* (0.140)	0.242*** (0.088)	0.064* (0.033)	0.065** (0.027)
Female			-0.428*** (0.080)	-0.111*** (0.020)	-0.080*** (0.029)
General altruism				-0.197** (0.089)	-0.141** (0.060)
Risk aversion					-0.086*** (0.031)
Time discounting					-0.177** (0.078)
Trust					0.073 (0.046)
Negative reciprocity					0.058 (0.045)
Positive reciprocity					0.034 (0.024)
Non-medical students		0.130 (0.089)	0.095 (0.075)	0.046 (0.028)	0.040* (0.021)
<i>μ</i>					
Constant	0.770*** (0.023)	0.839*** (0.040)	0.720*** (0.045)	0.601*** (0.048)	0.627*** (0.112)
2ND cohort		-0.001 (0.004)	-0.00004 (0.005)	0.005 (0.014)	-0.006 (0.023)
3RD cohort		-0.013*** (0.003)	-0.018*** (0.006)	-0.048*** (0.016)	-0.095** (0.043)
4TH cohort		-0.006 (0.005)	-0.010 (0.008)	-0.013 (0.020)	-0.053 (0.039)
Female			0.019*** (0.006)	0.040*** (0.014)	0.062* (0.034)
General altruism				0.117*** (0.041)	0.209** (0.106)
Risk aversion					0.034 (0.066)
Time discounting					0.273** (0.117)
Trust					-0.069 (0.043)
Negative reciprocity					-0.041 (0.065)
Positive reciprocity					-0.022 (0.056)
Non-medical students		-0.0004 (0.004)	-0.0004 (0.006)	-0.002 (0.017)	-0.018 (0.030)
<i>N</i>	878	878	878	873	873
Log-likelihood	-6,615.2	-6,631.76	-6,624.59	-6,582.66	-6,578.67
<i>Notes.</i>				* p<0.10; ** p<0.05; *** p<0.01	

D.5 Alternative behavioral model: Fehr and Schmidt (1999)

D.5.1 Fehr and Schmidt parametric form for the utility function

For an alternative variant for the utility function u_i , we now consider a Fehr and Schmidt parametric form defined as:

$$u_i(s, o, \alpha_i, \beta_i) = s - \alpha_i \max(o - s, 0) - \beta_i \max(s - o, 0)$$

where α_i represents the penalty implied by the difference between other and self and β_i represents the penalty implied by the difference between self and other, for participant i . $\alpha_i > 0$ corresponds to aversion to disadvantageous inequality (and $\alpha_i < 0$ corresponds to a taste to disadvantageous inequality, being behindness averse), $\beta_i > 0$ corresponds to aversion to advantageous inequality (aheadness aversion).

For the sake of notational simplicity, we denote $\theta_i = (\alpha_i, \beta_i, \mu_i)$ for the Fehr-Schmidt preference functional. Similarly to the estimation procedure for the CES preference functional form, the estimation procedure for the Fehr-Schmidt preference functional form accounts for parameter constraints on noise with an exponential transformation ($\mu_i = g^\mu(\zeta_i^\mu) = \exp(\zeta_i^\mu)$) of the parameter value to guarantee the noise parameter is positive. Contrary to the CES preference function, no theoretical restrictions are put on the parameters α_i and β_i for the Fehr-Schmidt preference functional and transformation functions correspond to the identity function $id()$ on \mathbb{R} . In vector notation, for the Fehr-Schmidt preference functional, $\theta_i = g(\zeta_i)$ with $\zeta_i = (\alpha_i, \beta_i, \zeta_i^\mu)$ and $g() = (id(), id(), g^\mu())$.

The regression tables report both the regression coefficients ζ_i and the value of the transformed regression coefficients in noise ,*i.e.* when regression coefficients are transformed back to the original scale.

D.5.2 Aggregate estimation results for Fehr and Schmidt preferences

Table D.5.1 shows the estimation results with transformation of the estimated parameters into preference parameters, noise, and marginal effects. The base value of parameter α in

Model (1) shows a taste for disadvantageous inequality. Aversion to disadvantageous inequality increases with term, with a maximum attained for clinical studies. This effect remains stable after controlling for gender, altruism, preferences and personality measures. Altruist-oriented subjects display a lower aversion to disadvantageous inequality (or higher taste for disadvantageous inequality). The same hold for trust, positive and negative reciprocity. On the other hand, gender and discounting are associated to higher aversion to disadvantageous inequality. The base Model (1) shows a high aversion to advantageous inequality, as measure by parameter β . Parameter β decreases with term, with a minimum attained for clinical studies. On the opposite, gender and altruism increase aversion to advantageous inequality. The same applies for risk aversion, discounting and positive reciprocity. Noise tends to be lower for pre-clinical students. Positive reciprocity and discounting is found to increase noise.

Table D.5.1: Aggregate estimations, preference parameters, noise, and marginal effects, Fehr and Schmidt preferences

Model	(1)	(2)	(3)	(4)	(5)	(6)
α						
Constant	-0.260*** (0.013)	-0.346*** (0.024)	-0.388*** (0.028)	-0.264*** (0.041)	-0.092*** (0.090)	-0.105*** (0.093)
Pre-clinical		0.108*** (0.032)	0.117*** (0.032)	0.105*** (0.033)	0.078*** (0.032)	0.076*** (0.033)
Clinical		0.164*** (0.037)	0.170*** (0.039)	0.136*** (0.037)	0.078*** (0.040)	0.075*** (0.041)
Practical Year		0.103*** (0.049)	0.122*** (0.051)	0.120*** (0.056)	0.106*** (0.071)	0.106*** (0.075)
Female			0.068*** (0.029)	0.084*** (0.029)	0.089*** (0.029)	0.075*** (0.030)
General altruism				-0.033*** (0.009)	-0.019*** (0.008)	-0.020*** (0.009)
Risk aversion					-0.006 (0.010)	-0.009** (0.011)
Time discounting					0.041*** (0.014)	0.043*** (0.014)
Trust					-0.035*** (0.007)	-0.036*** (0.007)
Negative reciprocity					-0.011*** (0.010)	-0.009** (0.010)
Positive reciprocity					-0.011*** (0.009)	-0.010*** (0.009)
Emotionality						0.011*** (0.008)
Extraversion						0.006** (0.008)
Agreeableness						-0.0002 (0.008)
Conscientiousness						0.005 (0.008)
Openness						-0.004 (0.007)
β						
Constant	0.970*** (0.018)	1.105*** (0.042)	0.954*** (0.044)	0.687*** (0.060)	0.557*** (0.112)	0.493*** (0.137)
Pre-clinical		-0.167*** (0.051)	-0.142*** (0.052)	-0.114*** (0.055)	-0.093*** (0.055)	-0.108*** (0.064)
Clinical		-0.230*** (0.052)	-0.223*** (0.052)	-0.184*** (0.055)	-0.168*** (0.057)	-0.192*** (0.064)
Practical Year		-0.196*** (0.067)	-0.138*** (0.075)	-0.122*** (0.087)	-0.050 (0.109)	-0.060 (0.119)
Female			0.250*** (0.045)	0.168*** (0.041)	0.158*** (0.040)	0.149*** (0.052)
General altruism				0.081*** (0.013)	0.075*** (0.014)	0.078*** (0.014)
Risk aversion					0.020*** (0.015)	0.015* (0.018)
Time discounting					0.023*** (0.015)	0.028*** (0.015)
Trust					-0.003 (0.009)	-0.001 (0.010)
Negative reciprocity					-0.011** (0.016)	-0.008 (0.018)
Positive reciprocity					0.031*** (0.015)	0.033*** (0.016)
Emotionality						0.007* (0.011)
Extraversion						-0.002 (0.012)
Agreeableness						-0.011** (0.016)
Conscientiousness						0.023*** (0.015)
Openness						-0.002 (0.009)
μ						
Constant	2.554*** (0.098)	2.714*** (0.178)	2.342*** (0.213)	1.846*** (0.292)	1.714*** (0.511)	1.578*** (0.535)
Pre-clinical		-0.478*** (0.234)	-0.316** (0.238)	-0.200** (0.224)	-0.149* (0.210)	-0.228*** (0.213)
Clinical		-0.282 (0.272)	-0.181 (0.272)	-0.142 (0.259)	-0.163* (0.257)	-0.206** (0.255)
Practical Year		-0.015 (0.348)	0.332 (0.465)	0.310* (0.473)	0.750*** (0.665)	0.624*** (0.648)
Female			0.493*** (0.292)	0.147* (0.213)	0.164** (0.196)	0.189*** (0.225)
General altruism				0.129*** (0.044)	0.115*** (0.063)	0.100*** (0.065)
Risk aversion					-0.006 (0.062)	-0.027 (0.069)
Time discounting					0.079*** (0.064)	0.085*** (0.059)
Trust					-0.045*** (0.047)	-0.041*** (0.047)
Negative reciprocity					-0.044** (0.069)	-0.019 (0.069)
Positive reciprocity					0.067*** (0.060)	0.076*** (0.058)
Emotionality						-0.002 (0.042)
Extraversion						-0.045*** (0.046)
Agreeableness						0.004 (0.059)
Conscientiousness						0.042** (0.054)
Openness						-0.010 (0.036)
N	733	733	733	729	729	729
Log-likelihood	-13,148.31	-12,815.62	-12,685.95	-12,176.39	-11,797.29	-11,764.14
Notes.	* p<0.10; ** p<0.05; *** p<0.01					

D.5.3 Random coefficient model for Fehr and Schmidt preferences

Table D.5.2: Random coefficient model, covariance parameter estimates, Fehr and Schmidt preferences

Model:	(1)	(2)	(3)	(4)	(5)	(6)
$\Omega_{[a,a]}$	0.127*** (0.010)	0.124*** (0.010)	0.123*** (0.010)	0.120*** (0.010)	0.109*** (0.008)	0.109*** (0.009)
$\Omega_{[r,r]}$	0.272*** (0.028)	0.256*** (0.027)	0.234*** (0.025)	0.207*** (0.022)	0.204*** (0.022)	0.205*** (0.022)
$\Omega_{[\mu,\mu]}$	0.530*** (0.063)	0.536*** (0.065)	0.516*** (0.063)	0.505*** (0.062)	0.461*** (0.057)	0.465*** (0.058)
$\Omega_{[a,r]}$	-0.044*** (0.011)	-0.037*** (0.011)	-0.040*** (0.011)	-0.031*** (0.010)	-0.028*** (0.009)	-0.028*** (0.009)
$\Omega_{[a,\mu]}$	0.124*** (0.020)	0.129*** (0.019)	0.122*** (0.019)	0.125*** (0.019)	0.106*** (0.016)	0.103*** (0.016)
$\Omega_{[r,\mu]}$	0.113*** (0.025)	0.105*** (0.026)	0.088*** (0.023)	0.078*** (0.023)	0.076*** (0.022)	0.081*** (0.024)
N	733	733	733	729	729	705
Log-likelihood	-4,755.7	-4,755.29	-4,757.09	-4,738.65	-4,743.42	-4,565.67
Notes.	*p<0.10; **p<0.05; ***p<0.01					

Table D.5.3 shows the estimation results with transformation of the estimated parameters. The base value of parameter α shows a taste for disadvantagenous inequality. Aversion to disadvantagenous inequality increases with term, with a maximum attained for clinical studies. This effect remains stable after controlling for gender, altruism, but vanish when controlling for preferences and personnality measures for clinical and practical. Altruist-oriented subjects have lower aversion to disadvantagenous inequality (or higher taste for disadvantagenous inequality). The same hold for trust, positive reciprocity. Negative reciprocity is no longer significant. On the other hand, gender and discounting are associated to higher aversion to disadvantagenous inequality.

Parameter β decreases with term, with a minimum attains for clinical studies. On the opposite, gender, altruism, risk aversion and positive reciprocity increase aversion to advantagenous inequality. The impact of negative reciprocity is no longer significant. Noise tends to be lower for pre-clinical students. Gender, altruism, positive reciprocity and discounting are found to increase noise.

Figure D.5.1 shows the distribution of observed heterogeneity implied by the different set of covariates and Figure D.5.2 shows the corresponding distribution of unobserved heterogeneity.

Table D.5.3: Random coefficient model, preference parameters, Fehr and Schmidt preferences

Model:	(1)	(2)	(3)	(4)	(5)	(6)
α						
Constant	-0.249*** (0.015)	-0.338*** (0.023)	-0.363*** (0.026)	-0.214*** (0.040)	-0.060 (0.067)	-0.085* (0.045)
Pre-clinical		0.134*** (0.034)	0.111*** (0.031)	0.111*** (0.034)	0.070*** (0.025)	0.095*** (0.034)
Clinical		0.164*** (0.036)	0.144*** (0.034)	0.127*** (0.033)	0.070** (0.033)	0.063* (0.037)
Practical year		0.113** (0.048)	0.084* (0.047)	0.066 (0.046)	0.053 (0.038)	0.041 (0.048)
Female			0.064** (0.025)	0.095*** (0.029)	0.102*** (0.028)	0.080*** (0.026)
General altruism				-0.417*** (0.083)	-0.199*** (0.059)	-0.203*** (0.073)
Risk aversion					0.025 (0.078)	0.042 (0.072)
Time discounting					0.281*** (0.087)	0.245*** (0.079)
Trust					-0.314*** (0.058)	-0.258*** (0.035)
Negative reciprocity					-0.060 (0.072)	-0.087 (0.072)
Positive reciprocity					-0.185** (0.081)	-0.136*** (0.052)
Emotionality						0.044 (0.031)
Extraversion						0.063 (0.041)
Agreeableness						-0.073** (0.036)
Conscientiousness						-0.016 (0.034)
Openness						-0.006 (0.030)
β						
Constant	1.098*** (0.026)	1.271*** (0.041)	1.127*** (0.048)	0.761*** (0.049)	0.711*** (0.090)	0.707*** (0.166)
Pre-clinical		-0.188*** (0.051)	-0.196*** (0.046)	-0.154*** (0.042)	-0.150*** (0.048)	-0.144*** (0.039)
Clinical		-0.377*** (0.057)	-0.376*** (0.051)	-0.318*** (0.045)	-0.313*** (0.052)	-0.297*** (0.062)
Practical year		-0.280*** (0.083)	-0.308*** (0.078)	-0.239*** (0.076)	-0.280*** (0.053)	-0.259*** (0.084)
Female			0.247*** (0.039)	0.206*** (0.044)	0.182*** (0.047)	0.162*** (0.042)
General altruism				0.936*** (0.128)	0.795*** (0.136)	0.601*** (0.190)
Risk aversion					0.229*** (0.080)	0.167* (0.089)
Time discounting					0.047 (0.097)	-0.159** (0.066)
Trust					0.070 (0.066)	0.063 (0.090)
Negative reciprocity					-0.124 (0.076)	0.012 (0.141)
Positive reciprocity					0.237*** (0.087)	0.362*** (0.090)
Emotion						0.082 (0.059)
Extraversion						0.033 (0.047)
Agreeableness						-0.018 (0.046)
Conscientiousness						0.049 (0.048)
Openness						0.017 (0.052)
μ						
Constant	0.868*** (0.035)	0.888*** (0.056)	0.804*** (0.067)	0.677*** (0.047)	0.862*** (0.115)	0.661*** (0.098)
Pre-clinical		0.012 (0.013)	0.004 (0.010)	0.014 (0.013)	-0.002 (0.012)	0.020 (0.014)
Clinical		-0.025** (0.012)	-0.031** (0.013)	-0.028** (0.012)	-0.057*** (0.018)	-0.036*** (0.012)
Practical year		-0.012 (0.015)	-0.024 (0.020)	-0.023 (0.018)	-0.035*** (0.011)	-0.007 (0.013)
Female			0.032*** (0.009)	0.042*** (0.012)	0.050*** (0.014)	0.041** (0.016)
General altruism				0.055*** (0.020)	0.129*** (0.040)	0.055** (0.023)
Risk aversion					-0.020 (0.023)	-0.034** (0.017)
Time discounting					0.121** (0.055)	0.098* (0.050)
Trust					-0.079*** (0.019)	-0.024 (0.021)
Negative reciprocity					-0.021 (0.018)	-0.012 (0.011)
Positive reciprocity					-0.051* (0.027)	0.038 (0.028)
Emotionality						0.043*** (0.013)
Extraversion						0.023 (0.015)
Agreeableness						-0.019* (0.010)
Conscientiousness						-0.010 (0.009)
Openness						-0.022 (0.018)
N	733	733	733	729	729	705
Log-likelihood	-4,755.7	-4,755.29	-4,757.09	-4,738.65	-4,743.42	-4,565.67

Notes.

*p<0.10; **p<0.05; ***p<0.01

Figure D.5.1: Distributions of parameters a , r and noise based on observed heterogeneity for the random coefficient model with different sets of covariates, Fehr and Schmidt preferences

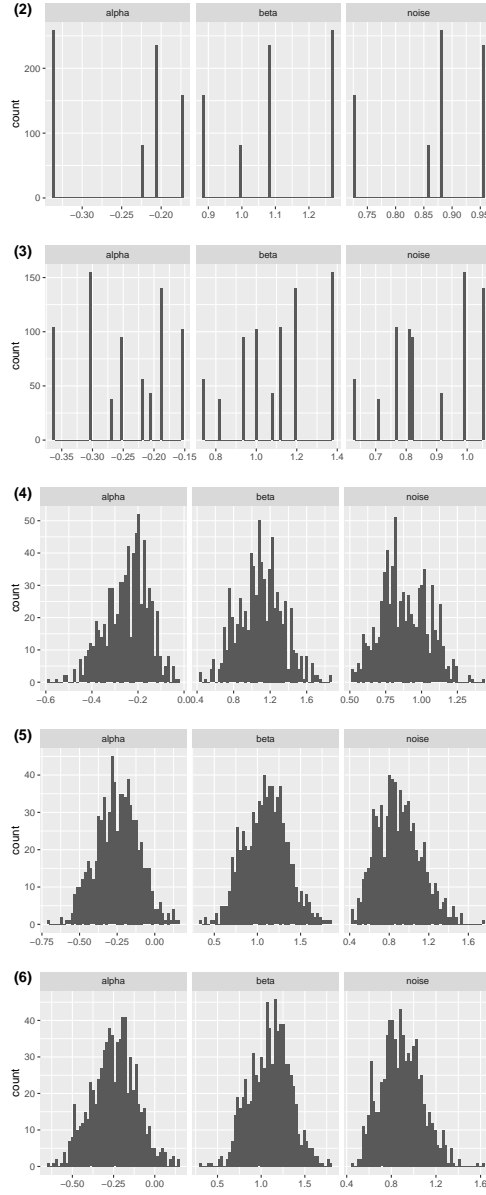
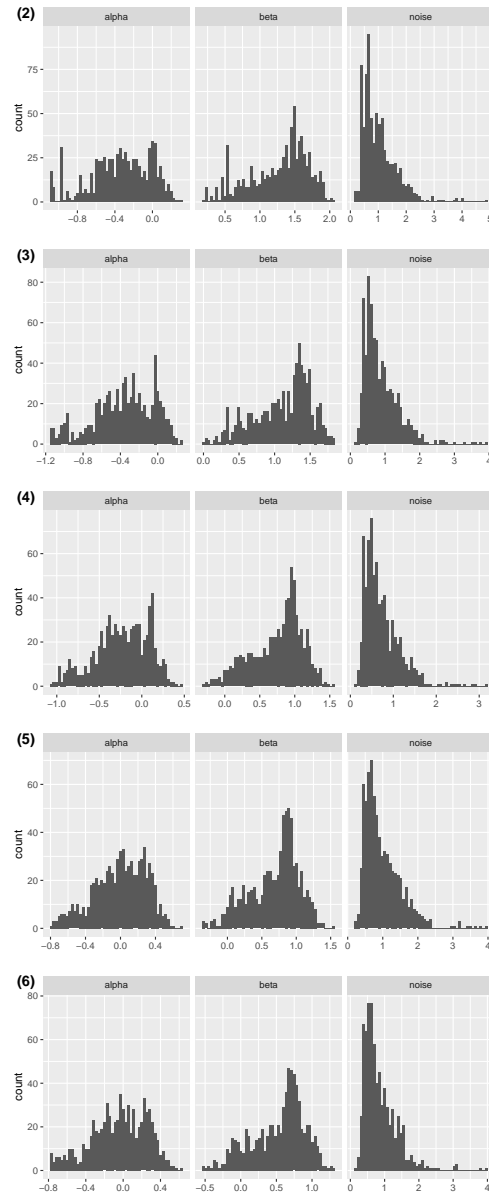


Figure D.5.2: Distributions of parameters a , r and noise based on unobserved heterogeneity for the random coefficient model with different sets of covariates, Fehr and Schmidt preferences



D.6 Estimation results for patient-regarding altruism, specialty choices, and income expectations

Tables D.6.1 shows the estimation results for aggregate estimations and Table D.6.2 shows the result from the random coefficient model including medical students' expectations on their future income, as reported in the main text. The components of the Ω matrix are given in a separate Table D.6.3.

Table D.6.4 shows the most preferred stated specialty choices crossed with study cohorts. Table D.6.5 shows the estimation results for aggregate estimations and Table D.6.6 shows the result from the random coefficient model including the four most preferred specialties, as reported in the main text. The components of the Ω matrix are given in a separate Table D.6.7. Figure D.6.1 shows the typical indifference curves for the different specialty choice based on preference parameters from Model (1).

Table D.6.1: Aggregate estimations including income expectations, preference parameters a and r , marginal effects, CES preferences

Model	(1)	(2)	(3)	(4)
a				
Constant	0.266*** (0.022)	0.336*** (0.030)	0.250*** (0.034)	0.798*** (0.112)
Expected income	0.139*** (0.030)	0.132*** (0.035)	0.107*** (0.033)	0.040*** (0.021)
Pre-clinical			0.128*** (0.043)	0.073*** (0.053)
Clinical			0.218*** (0.043)	0.094*** (0.057)
Practical year			0.110*** (0.103)	0.082*** (0.055)
Female		-0.106*** (0.025)	-0.095*** (0.026)	-0.035*** (0.038)
General altruism				-0.041*** (0.025)
Risk aversion				-0.009*** (0.012)
Discounting				0.006*** (0.009)
Trust				-0.019*** (0.006)
Negative reciprocity				-0.006*** (0.012)
Positive reciprocity				-0.026*** (0.011)
Emotionality				0.004** (0.006)
Extraversion				0.0004 (0.011)
Agreeableness				0.013*** (0.010)
Conscientiousness				-0.018*** (0.010)
Openness				-0.003* (0.007)
r				
Constant	-1.215*** (0.124)	-0.512*** (0.142)	-0.499*** (0.210)	0.081 (0.393)
Expected income	0.485*** (0.177)	0.273*** (0.159)	0.211*** (0.170)	0.040 (0.115)
Pre-clinical			0.001 (0.224)	0.076* (0.168)
Clinical			0.027 (0.350)	0.149*** (0.195)
Practical year			-0.348 (0.769)	0.183*** (0.245)
Female		-1.107*** (0.226)	-1.193*** (0.257)	-0.347*** (0.256)
General altruism				-0.049*** (0.066)
Risk aversion				0.011 (0.048)
Discounting				-0.069*** (0.056)
Trust				0.014 (0.030)
Negative reciprocity				0.002 (0.058)
Positive reciprocity				-0.058*** (0.048)
Emotionality				-0.002 (0.025)
Extraversion				-0.006 (0.049)
Agreeableness				0.053*** (0.044)
Conscientiousness				-0.082*** (0.047)
Openness				0.0005 (0.028)
μ				
Constant	2.548*** (0.121)	2.390*** (0.169)	2.088*** (0.216)	1.441*** (0.371)
Expected income	0.001 (0.182)	-0.035 (0.206)	0.055 (0.201)	0.201*** (0.179)
Pre-clinical			0.009 (0.226)	-0.123** (0.185)
Clinical			0.190 (0.347)	-0.052 (0.212)
Practical year			1.042*** (0.779)	0.151 (0.372)
Female		0.151 (0.208)	0.277*** (0.216)	0.070 (0.202)
General altruism				-0.005 (0.050)
Risk aversion				-0.033* (0.051)
Discounting				0.026 (0.055)
Trust				0.011 (0.032)
Negative reciprocity				0.014 (0.055)
Positive reciprocity				0.080*** (0.057)
Emotionality				-0.007 (0.034)
Extraversion				-0.041*** (0.046)
Agreeableness				-0.034** (0.058)
Conscientiousness				0.026* (0.040)
Openness				-0.011 (0.032)
N	693	693	693	693
Log-likelihood	-12,467.94	-12,384.92	-12,048.49	-11,276.86
Notes.	*p<0.10; **p<0.05; ***p<0.01			

Table D.6.2: Random coefficient model including income expectations, preference parameters a and r , marginal effects, CES preferences

Model	(1)	(2)	(3)	(4)
a				
Constant	0.268*** (0.021)	0.316*** (0.033)	0.220*** (0.031)	0.618*** (0.085)
Expected income	0.018*** (0.005)	0.018*** (0.006)	0.012*** (0.004)	0.015*** (0.006)
Pre-clinical			0.134*** (0.032)	0.109*** (0.033)
Clinical			0.249*** (0.041)	0.155*** (0.039)
Practical year			0.205*** (0.061)	0.146*** (0.045)
Female		-0.067** (0.030)	-0.058*** (0.021)	-0.021 (0.019)
General altruism				-0.044*** (0.010)
Risk aversion				-0.011*** (0.004)
Discounting				0.012*** (0.004)
Trust				-0.021*** (0.004)
Negative reciprocity				-0.008** (0.003)
Positive reciprocity				-0.017*** (0.004)
Emotionality				0.0004 (0.007)
Extraversion				0.001 (0.007)
Agreeableness				-0.001 (0.007)
Conscientiousness				-0.011** (0.005)
Openness				-0.003 (0.007)
r				
Constant	-1.024*** (0.166)	-0.368*** (0.138)	-0.597*** (0.148)	-0.105 (0.244)
Expected income	0.102 (0.065)	0.049 (0.046)	0.115 (0.079)	0.024 (0.017)
Pre-clinical			0.015 (0.081)	-0.006 (0.017)
Clinical			0.179* (0.105)	0.028 (0.019)
Practical year			0.294** (0.128)	0.056* (0.030)
Female		-0.264*** (0.056)	-0.392*** (0.096)	-0.058*** (0.021)
General altruism				-0.091 (0.055)
Risk aversion				-0.006 (0.033)
Discounting				-0.093*** (0.035)
Trust				0.037 (0.031)
Negative reciprocity				-0.024 (0.032)
Positive reciprocity				-0.007 (0.023)
Emotionality				-0.031 (0.033)
Extraversion				-0.024 (0.024)
Agreeableness				0.012 (0.020)
Conscientiousness				-0.007 (0.018)
Openness				0.011 (0.015)
μ				
Constant	0.775*** (0.037)	0.668*** (0.042)	0.742*** (0.048)	0.587*** (0.116)
Expected income	-0.002 (0.005)	-0.0001 (0.005)	-0.001 (0.004)	-0.006 (0.025)
Pre-clinical			-0.002 (0.004)	0.006 (0.029)
Clinical			-0.015*** (0.005)	-0.077** (0.037)
Practical year			-0.008 (0.006)	-0.051 (0.034)
Female		0.022*** (0.008)	0.012** (0.005)	0.070* (0.038)
General altruism				0.173 (0.123)
Risk aversion				-0.059 (0.073)
Discounting				0.155** (0.077)
Trust				-0.030 (0.058)
Negative reciprocity				0.021 (0.055)
Positive reciprocity				0.045 (0.066)
Emotionality				0.043 (0.030)
Extraversion				-0.004 (0.035)
Agreeableness				-0.011 (0.032)
Conscientiousness				-0.010 (0.026)
Openness				-0.044 (0.027)
N	693	693	693	693
Log-likelihood	-5,235.95	-5,230.15	-5,227.93	-5,216.73
Notes.	*p<0.10; **p<0.05; ***p<0.01			

Table D.6.3: Random coefficient model including income expectations, covariance parameter estimates, CES preferences

Model	(1)	(2)	(3)	(4)
$\Omega_{[a,a]}$	1.964*** (0.199)	1.921*** (0.191)	1.726*** (0.174)	1.360*** (0.135)
$\Omega_{[r,r]}$	1.397*** (0.166)	1.255*** (0.150)	1.233*** (0.149)	1.155*** (0.138)
$\Omega_{[\mu,\mu]}$	0.373*** (0.039)	0.358*** (0.041)	0.343*** (0.041)	0.332*** (0.041)
$\Omega_{[a,r]}$	0.453*** (0.136)	0.363*** (0.118)	0.329*** (0.110)	0.281*** (0.089)
$\Omega_{[a,\mu]}$	-0.237*** (0.070)	-0.191*** (0.067)	-0.151*** (0.057)	-0.133** (0.056)
$\Omega_{[r,\mu]}$	-0.546*** (0.065)	-0.485*** (0.063)	-0.463*** (0.064)	-0.425*** (0.061)
N	693	693	693	693
Log-likelihood	-5,235.95	-5,230.15	-5,227.93	-5,216.73
Notes.	* p<0.10; ** p<0.05; *** p<0.01			

Table D.6.4: Descriptive statistics on stated specialty choice

	Freshman	Pre-clinical	Clinical	Practical Year	Overall
Surgery	75 (29.0%)	42 (17.9%)	15 (9.5%)	5 (6.2%)	137 (18.7%)
Internal medicine	25 (9.7%)	34 (14.5%)	30 (19.0%)	21 (25.9%)	110 (15.0%)
Pediatrics	37 (14.3%)	29 (12.3%)	22 (13.9%)	9 (11.1%)	97 (13.2%)
Neurology/psychiatry	34 (13.1%)	32 (13.6%)	13 (8.2%)	5 (6.2%)	84 (11.5%)
Anesthesia	12 (4.6%)	15 (6.4%)	15 (9.5%)	8 (9.9%)	50 (6.8%)
Others	22 (8.5%)	23 (9.8%)	9 (5.7%)	5 (6.2%)	59 (8.1%)
Orthopedics	16 (6.2%)	15 (6.4%)	17 (10.8%)	2 (2.5%)	50 (6.8%)
General medicine	11 (4.2%)	18 (7.7%)	11 (7.0%)	6 (7.4%)	46 (6.3%)
Gynecology	9 (3.5%)	7 (3.0%)	5 (3.2%)	5 (6.2%)	26 (3.5%)
Radiology/nuclear medicine	8 (3.1%)	3 (1.3%)	1 (0.6%)	6 (7.4%)	18 (2.5%)
Ophthalmology	0 (0.0%)	2 (0.9%)	7 (4.4%)	3 (3.7%)	12 (1.6%)
Dermatology	2 (0.8%)	2 (0.9%)	5 (3.2%)	3 (3.7%)	12 (1.6%)
Forensic medicine	3 (1.2%)	4 (1.7%)	2 (1.3%)	1 (1.2%)	10 (1.4%)
Otorhinolaryngology	2 (0.8%)	3 (1.3%)	3 (1.9%)	0 (0.0%)	8 (1.1%)
Urology	1 (0.4%)	3 (1.3%)	2 (1.3%)	2 (2.5%)	8 (1.1%)
Dentistry and maxillary Surgery	0 (0.0%)	2 (0.9%)	1 (0.6%)	0 (0.0%)	3 (0.4%)
Laboratory medicine	2 (0.8%)	1 (0.4%)	0 (0.0%)	0 (0.0%)	3 (0.4%)
N	259 (100%)	235 (100%)	158 (100%)	81 (100%)	733 (100%)

Notes. This table shows the most preferred stated specialty choices crossed with study cohorts. Specialties sorted in descending order.

Table D.6.5: Aggregate estimations including specialty choice, preference parameters, noise, and marginal effects, CES preferences

Model	(1)	(2)	(3)	(4)
α				
Constant	0.401*** (0.022)	0.448*** (0.029)	0.323*** (0.062)	0.757*** (0.124)
Surgery	-0.120*** (0.039)	-0.076*** (0.054)	-0.025** (0.041)	-0.027** (0.050)
Internal Medicine	-0.070*** (0.045)	-0.007 (0.049)	-0.029** (0.047)	-0.010 (0.031)
Pediatrics	-0.138*** (0.044)	-0.047*** (0.059)	-0.034** (0.061)	-0.032*** (0.041)
Neurology/Psychiatry	-0.103*** (0.062)	-0.071*** (0.080)	-0.031** (0.054)	-0.020* (0.052)
Pre-clinical			0.156*** (0.074)	0.083*** (0.048)
Clinical			0.252*** (0.074)	0.113*** (0.051)
Practical year			0.121*** (0.122)	0.062*** (0.071)
Female		-0.144*** (0.032)	-0.130*** (0.056)	-0.049*** (0.041)
General altruism				-0.053*** (0.021)
Risk aversion				-0.013*** (0.013)
Discounting				0.010*** (0.009)
Trust				-0.018*** (0.008)
Negative reciprocity				0.002 (0.018)
Positive reciprocity				-0.024*** (0.011)
Emotionality				0.003 (0.008)
Extraversion				0.006*** (0.012)
Agreeableness				0.007** (0.013)
Conscientiousness				-0.010*** (0.015)
Openness				0.0005 (0.007)
τ				
Constant	-0.820*** (0.155)	-0.492*** (0.178)	-0.581*** (0.356)	-0.462*** (0.613)
Surgery	-0.311 (0.228)	0.131 (0.291)	0.195** (0.220)	0.069 (0.215)
Internal Medicine	-0.284 (0.289)	0.253*** (0.268)	0.189* (0.294)	0.201*** (0.187)
Pediatrics	-0.219 (0.266)	0.395*** (0.272)	0.422*** (0.269)	0.298*** (0.267)
Neurology/Psychiatry	-0.199 (0.317)	0.133 (0.396)	0.282*** (0.288)	0.317*** (0.272)
Female		-1.231*** (0.374)	-1.387*** (0.536)	-0.489*** (0.335)
General altruism				-0.129*** (0.098)
Pre-clinical			0.099 (0.390)	0.145** (0.263)
Clinical			0.082 (0.562)	0.307*** (0.256)
Practical year			-0.167 (0.710)	-0.084 (0.725)
Risk aversion				-0.031* (0.070)
Discounting				-0.090*** (0.086)
Trust				0.055*** (0.052)
Negative reciprocity				0.050** (0.128)
Positive reciprocity				-0.062*** (0.062)
Emotionality				-0.024* (0.051)
Extraversion				0.020 (0.072)
Agreeableness				0.031* (0.074)
Conscientiousness				-0.066*** (0.095)
Openness				0.016 (0.042)
μ				
constant	2.588*** (0.155)	2.747*** (0.255)	2.520*** (0.382)	2.157*** (0.624)
Surgery	-0.072 (0.241)	-0.440*** (0.332)	-0.437*** (0.282)	-0.254*** (0.243)
Internal Medicine	-0.052 (0.273)	-0.637*** (0.367)	-0.519*** (0.324)	-0.536*** (0.234)
Pediatrics	-0.282 (0.252)	-0.796*** (0.357)	-0.638*** (0.315)	-0.464*** (0.271)
Neurology/Psychiatry	0.598** (0.392)	0.180 (0.599)	0.012 (0.453)	-0.182 (0.312)
Female		0.113 (0.255)	0.235 (0.421)	0.013 (0.260)
General altruism				0.028 (0.064)
Pre-clinical			-0.176 (0.375)	-0.279*** (0.290)
Clinical			0.248 (0.529)	-0.159 (0.308)
Practical year			0.856*** (0.802)	0.834*** (0.910)
Risk aversion				-0.036 (0.085)
Discounting				0.044* (0.085)
Trust				-0.005 (0.050)
Negative reciprocity				-0.011 (0.097)
Positive reciprocity				0.053** (0.076)
Emotionality				-0.010 (0.047)
Extraversion				-0.084*** (0.073)
Agreeableness				0.001 (0.092)
Conscientiousness				0.047*** (0.077)
Openness				-0.020 (0.046)
N	733	733	733	729
Log-likelihood	-13,218.83	-13,129.38	-12,807.65	-11,931.48
Notes.	* p<0.10; ** p<0.05; *** p<0.01			

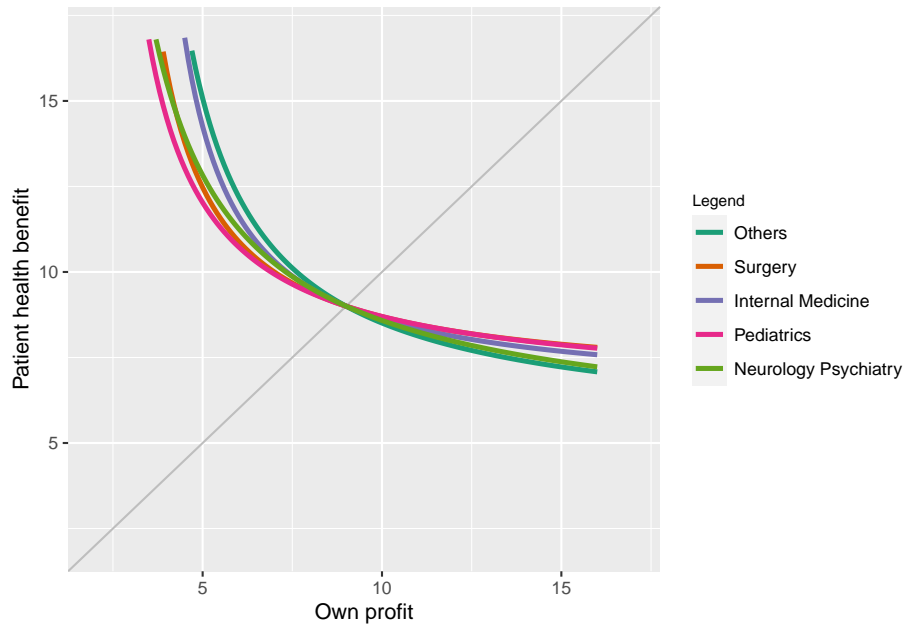
Table D.6.6: Random coefficient model including specialty choice, preference parameters, noise, and marginal effects, CES preferences

Model	(1)	(2)	(3)	(4)
<i>a</i>				
Constant	0.363*** (0.021)	0.426*** (0.028)	0.294*** (0.028)	0.612*** (0.055)
Surgery	-0.022*** (0.007)	-0.024** (0.010)	-0.010* (0.005)	-0.009* (0.005)
Internal Medicine	-0.012 (0.008)	-0.015* (0.008)	-0.012** (0.005)	-0.017 (0.011)
Pediatrics	-0.023*** (0.009)	-0.023** (0.009)	-0.015* (0.008)	-0.015*** (0.005)
Neurology/Psychiatry	-0.009 (0.008)	-0.008 (0.008)	0.001 (0.010)	-0.006 (0.008)
Pre-clinical			0.137*** (0.033)	0.110*** (0.020)
Clinical			0.253*** (0.039)	0.168*** (0.021)
Practical year			0.189*** (0.051)	0.135*** (0.024)
Female		-0.098*** (0.026)	-0.084*** (0.016)	-0.050** (0.022)
General altruism				-0.043*** (0.010)
Risk aversion				-0.018*** (0.007)
Time discounting				0.002 (0.009)
Trust				-0.020*** (0.005)
Negative reciprocity				0.011** (0.005)
Positive reciprocity				-0.013*** (0.005)
Emotionality				0.003 (0.006)
Extraversion				0.006 (0.006)
Agreeableness				-0.002 (0.008)
Conscientiousness				-0.015** (0.007)
Openness				-0.003 (0.005)
<i>r</i>				
Constant	-0.786*** (0.138)	-0.270** (0.136)	-0.405*** (0.150)	-0.658* (0.388)
Surgery	-0.038 (0.048)	-0.027 (0.044)	-0.019 (0.063)	-0.014 (0.019)
Internal Medicine	-0.056 (0.047)	-0.036 (0.046)	-0.080 (0.079)	-0.028 (0.029)
Pediatrics	-0.009 (0.061)	0.022 (0.047)	0.048 (0.075)	0.007 (0.015)
Neurology Psychiatry	0.003 (0.058)	0.039 (0.047)	0.107 (0.084)	0.015 (0.015)
Pre-clinical			-0.009 (0.056)	0.002 (0.015)
Clinical			0.143** (0.063)	0.038** (0.017)
Practical year			0.179** (0.083)	0.054*** (0.019)
Female		-0.170*** (0.036)	-0.315*** (0.059)	-0.074*** (0.026)
General altruism				-0.053 (0.042)
Risk aversion				-0.032 (0.028)
Time discounting				-0.138*** (0.047)
Trust				0.071** (0.033)
Negative reciprocity				0.035 (0.022)
Positive reciprocity				0.033 (0.024)
Emotionality				-0.018 (0.015)
Extraversion				-0.003 (0.016)
Agreeableness				-0.010 (0.021)
Conscientiousness				-0.013 (0.015)
Openness				0.003 (0.018)
<i>μ</i>				
Constant	0.782*** (0.038)	0.691*** (0.045)	0.741*** (0.049)	0.664*** (0.088)
Surgery	0.008 (0.010)	0.010 (0.013)	0.003 (0.006)	0.017 (0.025)
Internal Medicine	0.003 (0.010)	0.001 (0.014)	0.003 (0.007)	0.004 (0.029)
Pediatrics	-0.007 (0.012)	-0.016 (0.015)	-0.009 (0.008)	-0.035 (0.026)
Neurology Psychiatry	-0.023** (0.012)	-0.031* (0.017)	-0.020** (0.008)	-0.066* (0.035)
Pre-clinical			0.0004 (0.006)	-0.0001 (0.026)
Clinical			-0.020*** (0.007)	-0.086*** (0.032)
Practical year			-0.006 (0.008)	-0.024 (0.033)
Female		0.034*** (0.010)	0.019*** (0.006)	0.069** (0.030)
General altruism				0.152 (0.115)
Risk aversion				-0.050 (0.046)
Time discounting				0.244*** (0.085)
Trust				-0.036 (0.052)
Negative reciprocity				-0.043 (0.070)
Positive reciprocity				-0.018 (0.084)
Emotionality				0.029 (0.023)
Extraversion				-0.036 (0.028)
Agreeableness				0.009 (0.037)
Conscientiousness				-0.009 (0.024)
Openness				-0.027 (0.027)
<i>N</i>	733	733	733	729
Log-likelihood	-5,558.94	-5,550.56	-5,556.06	-5,522.58
Notes. * p<0.10; ** p<0.05; *** p<0.01				

Table D.6.7: Random coefficient model including specialty choice, covariance parameter estimates, CES preferences

Model	(1)	(2)	(3)	(4)
$\Omega_{[a,a]}$	2.010*** (0.193)	1.958*** (0.183)	1.809*** (0.169)	1.517*** (0.148)
$\Omega_{[r,r]}$	1.389*** (0.157)	1.251*** (0.141)	1.255*** (0.149)	1.186*** (0.137)
$\Omega_{[\mu,\mu]}$	0.368*** (0.041)	0.359*** (0.041)	0.348*** (0.039)	0.334*** (0.041)
$\Omega_{[a,r]}$	0.468*** (0.123)	0.376*** (0.112)	0.367*** (0.107)	0.339*** (0.094)
$\Omega_{[a,\mu]}$	-0.224*** (0.066)	-0.180*** (0.063)	-0.157*** (0.061)	-0.150*** (0.055)
$\Omega_{[r,\mu]}$	-0.536*** (0.066)	-0.484*** (0.062)	-0.471*** (0.063)	-0.436*** (0.059)
N	733	733	733	729
Log-likelihood	-5,558.94	-5,550.56	-5,556.06	-5,522.58
Notes.			* p<0.10; ** p<0.05; *** p<0.01	

Figure D.6.1: Indifference curves for different specialty choices based on random coefficient model, CES preferences



MONA GROSS

CONTACT DETAILS

Achterstr. 9
50678 Cologne, Germany
Email: mona.gross@wiso.uni-koeln.de

PERSONAL INFORMATION

Date of birth: 04 July 1991
Nationality: German

EDUCATION

- 10/2017–present **Doctoral studies in Business administration**
Faculty of Management, Economics and Social Sciences, University of Cologne
- 10/2014–04/2017 **Master of Science, Economics**, University of Cologne
- 09/2015–01/2016 Visiting Master student, Corvinus University of Budapest, Hungary
- 10/2010–11/2013 **Bachelor of Science Economics**, Heinrich-Heine University, Düsseldorf
- 08/2001–06/2010 **Abitur**, Gymnasium Koblenzer Straße, Düsseldorf
- 08/2007–06/2008 Visiting High School Student, O’Gorman High School, Sioux Falls, USA
-

PROFESSIONAL EXPERIENCE

- 09/2016–present **Research Assistant**, Dept. of Business Administration and Health Care Management, University of Cologne
- 06/2018–12/2018 **Assistant Editor**, Dr. Andreas Lehr, Agentur für Gesundheitspolitische Information
- 06/2016–02/2017 **Freelancer, Project Work ”Health Behavior and Medication Compliance”**, University of Cologne in cooperation with University of Paderborn
- 04/2016–06/2016 **Student Assistance, Project Work ”Managerial Risk Factors in Medicine”**, Dept. of Business Administration and Health Care Management, University of Cologne
- 10/2010–06/2016 **Student Assistance**, CETA Testsysteme GmbH, Depts. of Accounting and Purchasing, Hilden
- 04/2016–06/2016 **Intern (full time)**, Janssen-Cilag GmbH, Health Economics, Market Access & Governmental Affairs, Neuss
- 01/2014–04/2014 **Student Assistance**, Düsseldorf Institute for Competition Economics, Dept. of Health Economics, Heinrich-Heine University, Düsseldorf
-

PRICES AND PUBLICATIONS

- 2021 **The effects of audits and fines on upcoding in neonatology.** (with Hendrik Jürges and Daniel Wiesen) Health Economics.

- 2016 **The choice between a ritonavir-boosted protease inhibitor- and a non-nucleoside reverse transcriptase inhibitor-based regimen for initiation of antiretroviral treatment – results from an observational study in Germany.** (with Jörg Mahlich, Alexander Kuhlmann, Johannes Bogner, Hans Heiken, and Matthias Stoll) *Journal of Pharmaceutical Policy and Practice*, 9.
- 2016 **Unemployment, health, and education of HIV–infected males in Germany.** (with Annika Herr, Martin Hower, Alexander Kuhlmann, Jörg Mahlich, and Matthias Stoll) *International Journal of Public Health*, 61, 593-602.
- 2013 **Qualität von Krankenhäusern: Welche Maße werden in Deutschland diskutiert und was sagt die Praxis in den USA?**
Ordnungspolitische Perspektiven. Beiträge zum Wettbewerb im Krankenhaus- und Arzneimittelmarkt – Band 1: Krankenhäuser, 37, 21-45.
- 2013 **Price of the Düsseldorfer Chamber of Commerce and Industry for the best Bachelor thesis in Economics in 2013**

LANGUAGE SKILLS

German (first language), **English** (business fluent), **French** (advanced basic level)

TECHNICAL SKILLS

MS Office: Word, Excel, PowerPoint (excellent)

Statistical software: STATA (very good)

Others: Latex, ztree, SAP (good)

Cologne, 09 July 2021

Mona Groß

Eidesstattliche Erklärung

nach §8 Abs. 3 der Promotionsordnung vom 17.02.2015

Hiermit versichere ich an Eides Statt, dass ich die vorgelegte Arbeit selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Aussagen, Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Bei der Auswahl und Auswertung folgenden Materials haben mir die nachstehend aufgeführten Personen in der jeweils beschriebenen Weise entgeltlich/unentgeltlich (zutreffendes unterstreichen) geholfen:

Weitere Personen, neben den ggf. in der Einleitung der Arbeit aufgeführten Koautorinnen und Koautoren, waren an der inhaltlich-materiellen Erstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Ich versichere, dass ich nach bestem Wissen die reine Wahrheit gesagt und nichts verschwiegen habe.

Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Die Strafbarkeit einer falschen eidesstattlichen Versicherung ist mir bekannt, namentlich die Strafan drohung gemäß §156 StGB bis zu drei Jahren Freiheitsstrafe oder Geldstrafe bei vorsätzlicher Begehung der Tat bzw. gemäß §161 Abs. 1 StGB bis zu einem Jahr Freiheitsstrafe oder Geldstrafe bei fahrlässiger Begehung.

Köln, 09.07.2021

Ort, Datum



Unterschrift